Outils de spatialisation sonore pour terminaux mobiles : microphone 3D pour une utilisation nomade

Julian Palacino

OUTILS DE SPATIALISATION SONORE POUR TERMINAUX MOBILES

Microphone 3D pour une utilisation nomade

Thèse
présentée publiquement
à la Faculté des Sciences et Techniques de
l'Université du Maine

pour l'obtention du grade de Docteur en Sciences
par

Julian Palacino

acceptée sur proposition du jury :

Prof.	Alexander Raake,	rapporteur
Prof.	Gaël Richard,	rapporteur
Dr.	Marc Emerit	examinateur et membre du CST
Prof.	Philippe Herzog	examinateur et membre du CST
Dr.	Rozenn Nicol,	directeur de thèse
Prof.	Laurent Simon,	codirecteur de thèse
M.	Hervé Dejardin	invité

Le Mans, Université du Maine,
04 novembre 2014

Il n'existe personne qui aime la souffrance pour elle-même,
ni qui la recherche ni qui la veuille pour ce qu'elle est...
— Cicéron

à ma famille...

Remerciements

Tout d'abord, je tiens à remercier ma directrice de thèse Rozenn Nicol, de m'avoir donné l'opportunité de réaliser cette thèse à ses côtés et pour m'avoir encouragé et accompagné tout au long de cette agréable aventure. Je remercie également Laurent Simon, mon encadrant universitaire, pour ses conseils avisés et son soutien. Un grand merci également à Marc Emerit et Philippe Herzog, membres du CST, pour avoir suivi mes travaux avec un regard bienveillant et pour avoir répondu présent quand les échéances imposées approchaient.

Un grand merci également à Bruno Lozach, Stanislas Zimmermann et Christine Marcatté, responsables du laboratoire TPS et SVQ pour m'avoir apporté les moyens nécessaires pour mener à bien les trois années passées dans l'équipe. Je remercie également le LAUM pour m'avoir donné les moyens de participer aux conférences.

Je remercie aussi les membres de l'équipe TPS pour leur accueil et notamment Grégory Pallone et Julien Faure pour leur collaboration et le temps que nous avons passé à rédiger des brevets que malheureusement n'ont pas pu voir le jour et par les répétitions et les concerts avec les "Jambombers". A Arnaud Lefort et à Jérôme Magnoux pour leur intérêt dans mes travaux et le temps consacré à finaliser le prototype commencé par mon stagiaire Carl Meunier Njutsop que je remercie également. Merci à Noëlle, Serge et Pascal pour leur apport logistique facilitant le travail au quotidien.

A Jerôme, Stéphane et aux doctorants Sam, Joachim, Nirina et Romain ; aux post-docs Pyo, Sean et Magda et à l'apprenti Paul, qui ont fait de mon passage chez Orange une expérience très agréable et qui, avec leurs conversations, avaient la capacité de rendre plaisants les repas à la Sodhexo.

Merci à mon compagnon de rédaction Felipe et *mersi bras* à tous les Trégorrois qui font que leur belle région soit encore plus agréable et accueillante.

Je n'oublie pas mes parents et mon frère, puisque malgré l'éloignement, leur fierté me donne la force et l'envie d'aller plus loin.

Et pour finir, le plus grands des mercis à toi Caro pour m'avoir accompagné, encouragé, compris, chéri, engueulé (quand il fallait) et pour avoir eu le courage de relire et corriger toutes les fautes et coquilles de ce document.

Plestin-les-Grèves, août 2014 J.D.P.G.

Abstract

Mobile technologies (such as smartphones and tablets) are now common devices of the consumer market. In this PhD we want to use those technologies as the way to introduce tools of sound spatialization into the mass market. Today the size and the number of traducers used to pick-up and to render a spatial sound scene are the main factors which limit the portability of those devices.
As a first step, a listening test, based on a spatial audio recording of an opera, let us to evaluate the 3D audio technologies available today for headphone rendering. The results of this test show that, using the appropriate binaural decoding, it is possible to achieve a good binaural rendering using only the four sensors of the *Soundfield* microphone.
Then, the steps of the development of a 3D sound pick-up system are described. Several configurations are evaluated and compared. The device, composed of 3 cardioid microphones, was developed following an approach inspired by the sound source localization and by the concept of the "object format encoding". Using the microphone signals and an adapted post-processing it is possible to determine the directions of the sources and to extract a sound signal which is representative of the sound scene. In this way, it is possible to completely describe the sound scene and to compress the audio information. This method offer the advantage of being cross platform compatible. In fact, the sound scene encoded with this method can be rendered over any reproduction system.
A second method to extract the spatial information is proposed. It uses the real *in situ* characteristics of the microphone array to perform the sound scene analysis.
Some propositions are made to complement the 3D audio chain allowing to render the result of the sound scene encoding over a binaural system or any king of speaker array using all capabilities of the mobile devices.

Key words: spatial audio, HRTF, source localization, object format

Résumé

Les technologies nomades (smartphones, tablettes, ...) étant actuellement très répandues, nous avons souhaité, dans le cadre de cette thèse, les utiliser comme vecteur pour proposer au grand public des outils de spatialisation sonore. La taille et le nombre de transducteurs utilisés pour la captation et la restitution sonore spatialisée sont à ce jour la limitation principale pour une utilisation nomade. Dans une première étape, la captation d'un opéra pour une restitution sur des tablettes tactiles nous a permis d'évaluer les technologies audio 3D disponibles aujourd'hui. Les résultats de cette évaluation ont révélé que l'utilisation des quatre capteurs du microphone *Soundfield* donne de bons résultats à condition d'effectuer un décodage binaural adapté pour une restitution sur casque. Selon une approche inspirée des méthodes de localisation de source et le concept de format « objet », un prototype de prise de son 3D léger et compact a été développé. Le dispositif microphonique proposé se compose de trois capsules microphoniques cardioïdes. A partir des signaux microphoniques, un algorithme de post-traitement spatial est capable, d'une part, de déterminer la direction des sources et, d'autre part, d'extraire un signal sonore représentatif de la scène spatiale. Ces deux informations permettent ainsi de caractériser complètement la scène sonore 3D en fournissant un encodage spatial offrant le double avantage d'une compression de l'information audio et d'une flexibilité pour le choix du système de reproduction. En effet, la scène sonore ainsi encodée peut être restituée en utilisant un décodage adapté sur n'importe quel type de dispositif.
Plusieurs méthodes de localisation et différentes configurations microphoniques (géométrie et directivité) ont été étudiées.
Dans une seconde étape, l'algorithme d'extraction de l'information spatiale a été modifié pour prendre en compte les caractéristiques réelles *in situ* des microphones.
Des méthodes pour compléter la chaîne acoustique sont proposées permettant la restitution binaurale ainsi que sur tout autre dispositif de restitution. Elles proposent l'utilisation de capteurs de localisation présents sur les terminaux mobiles afin d'exploiter les capacités qu'ils offrent aujourd'hui.

Mots clefs : Son spatialisé, HRTF, localisation des sources, format objet

Table des matières

Table des figures

Index de tableaux

Glossaire

contralatéral Qui se trouve ou se produit du côté opposé du corps.

downmix Adapter des signaux sonores comportant un nombre signaux N à un dispositif ou une technique comportant un nombre N' de signaux, $N' < N$.

hand-traking Le *hand-traking* ou suivi de la main est un procédé utilisé lors de la captation d'une scène sonore spatialisée afin de compenser les mouvements des capteurs lors de la prise de son.

head-traking Le *head-traking* ou suivi de la tête est utilisé lors de la synthèse binaurale pour compenser les mouvements de la tête de l'auditeur au moment de la restitution sonore.

organisation tonotopique Chaque région de la membrane basilaire est sensible à une certaine plage de fréquences et peut être assimilée à un filtre passe-bande [Guillon, 2009].

ipsilatéral Qui se trouve ou se produit d'un même côté du corps.

MS Stéréo Methode de captation stéréophonique aussi connu comme *Mid-Side stereo* (US) *Middle-Side stereo* (UK) ou *Mitte-Scite stereo* (D).

plan horizontal C'est le plan perpendiculaire au Plmedian qui trouve origine sur la droite définie par les deux oreilles.

plan frontal Le plan frontal (ou coronal) est un plan vertical qui va d'un côté à l'autre du corps et le divise en partie antérieure et partie postérieure. Au niveau de la tête, il s'agit du plan vertical passant par la ligne définie par les deux oreilles.

plan médian C'est le plan qui divise la tête en deux dans le sens de la hauteur, la partageant en deux demi-têtes symétriques.

sweetspot Zone d'écoute idéale lors d'une restitution sur haut-parleurs.

upmix Adapter des signaux sonores comportant un nombre signaux N à un dispositif ou une technique comportant un nombre N' de signaux, $N' > N$. Les informations manquantes sont extraites des pistes sonores existantes.

Acronymes

ACP Analyse par Composantes Principales

ADSS Analyse De la Scène Sonore

ASA *Auditory Scene Analysis*

BRIR *Binaural Room Impulse Response*

CAN Convertisseur Analogique - Nunérique

cSCT *cumulative State Coherence Transform*

DAT *Digital Audio Tape*

DirAC *Directional Audio Coding*

ERB *Equivalent rectangular bandwidth*

FFT *Fast Fourier Transform Function*

GCC *Generalized Cross-Correlation*

HOA *Higher Order Ambisonics*

HPTF *Headphone Transfert Functions*

HRIR *Head Related Impulse Response*

HRTF *Head Related Transfert Function*

ICLD *Inter-channel Level Difference*

ICTD *Inter-channel Time Difference*

IID *Interaural Intensity Difference*

ILD *Interaural Level Difference*

IPD *Interaural Phase Difference*

ISD *Interaural Spectral Difference*

ISSD *Inter-Subject Spectral Difference*

ITD *Interaural Time Difference*

JND *Just-noticeable Difference*

LCRS *Left, Center, Right, Surround*

MAA *Minimum Audible Angle*

MAO Musique Assistée par Ordinateur

MCRA *Minima Controlled Recursive Averaging*

MDA *Multi Dimensional Audio*

MEMS *Micro-electro-mechanical Systems*

MIDI *Music Instrument Digital Interface*

MLP *Meridian Lossless Packing*

MLS *Maximum Length Sequence*

MUSHRA *MUltiple Stimuli with Hidden Reference and Anchor*

MUSIC *MUltiple Signal Classification*

MVDR *Minimum Variance Distortionless Response*

Ob-RTF *Object Related Transfert Function* ou fonction de transfert liée à l'objet

OSC *Open Sound Control*

PCM *Pulse-Code Modulation*

RSB Rapport Signal à Bruit

RSB Rapport Signal-à-Bruit

SAC *Spatial Audio Coding*

SAOC *Spatial Audio Object Coding*

SASC *Spatial Audio Scene Coding*

SDK *Software Developement Kit*

STPS *Spherical Thin Plate Spline*

TDOA *Time Difference Of Arrival*

TPS *Time Streched Pulse*

UIT Union Internationale des Télécomunications

VBAP *Vector Based Amplitude Panning*

VST *VirtualStudioTechnology*

WFS *Wave Field Synthesis*

Liste de symboles

a^* Désigne le complexe conjugué du nombre a.

$a * b$ Désigne le produit de convolution entre les vecteurs a et b.

c Célérité acoustique (340 ms^{-1}).

Δ Désigne une distance ou une différence entre deux valeurs.

$\bar{a}$ Désigne la valeur moyenne de a.

f Fréquence en Hz.

$\Im(a)$ Désigne la partie imaginaire du nombre a.

Désigne la fonction d'intercorrélation.

k Nombre d'onde ($\frac{\omega}{c}$).

$\| \vec{a} \|$ Désigne la norme du vecteur $\vec{a}$.

$\cdot$ Produit scalaire.

ϕ Désigne l'élévation en coordonnées sphériques (θ, ϕ, r).

$\Re(a)$ Désigne la partie réelle du nombre a.

θ Désigne l'azimut en coordonnées sphériques (θ, ϕ, r).

$|a|$ Désigne la valeur absolue du nombre a.

ω pulsation en radians par seconde $2\pi f$.

Introduction

Le son spatialisé, depuis les premiers pas de la stéréophonie, s'est heurté à des barrières technologiques limitant son développement et l'enfermant dans des laboratoires de recherche. Aujourd'hui, ces barrières se lèvent à grands pas grâce à l'évolution et à la miniaturisation des composants électroniques. Ces nouvelles technologies modifient également les modes de consommation et de création des médias. L'apparition du walkman dans les années 1980 et de l'iPod dans les années 2000, a marqué un tournant dans les usages faisant des dispositifs nomades le principal vecteur de diffusion et d'accès aux contenus multimédia.

Orange Labs, en tant que principal opérateur de téléphonie mobile en France, est devenu, grâce à l'élargissement des débits de transmission de données, l'un des principaux diffuseurs de contenus multimédia. Utilisant comme fer de lance son cœur de métier, cette entreprise a voulu démocratiser le son 3D auprès du grand public en utilisant les terminaux mobiles. Cette thèse a pour objectif d'implémenter dans les équipements nomades grand public des outils et des services exploitant l'audio spatialisé. Les outils développés pourront s'insérer dans l'ensemble de la chaîne audio, à savoir, lors de la captation, la production ou la synthèse, ainsi que dans la reproduction sonore en prenant en compte les contraintes imposées par ce type de dispositifs.

Nous disposons aujourd'hui d'un large éventail de technologies de spatialisation sonore, permettant de restituer l'ensemble des dimensions de l'espace sonore, aussi bien l'azimut, l'élévation que la profondeur. Un large choix de dispositifs microphoniques (couple stéréophonique, microphone *Soundfield*, microphone *Eigenmike*, têtes artificielles) s'offre pour la captation de toute ou partie de l'information spatiale. Pour la restitution, il existe également une grande variété de systèmes allant du simple casque d'écoute à des dispositifs multi haut-parleurs plus ou moins complexes (système 5.1, 10.2, 22.2). En général, tous ces équipements sont encombrants et requièrent des conditions de mise en œuvre très contraintes, ce qui les rend incompatibles avec le contexte nomade. De plus, dans tous les cas, les signaux issus des capteurs ou alimentant les haut-parleurs doivent être acheminés, traités ou stockés ce qui se traduit par une forte complexité du calcul qui doit être déployé. Enfin, même si les capacités d'analyse des dispositifs nomades augmentent de façon exponentielle, les traitements audio viennent directement impacter

la consommation énergétique qui est aussi un facteur déterminant de la portabilité d'une technologie pour son utilisation en mobilité.

La principale difficulté rencontrée dans cette étude a donc résidé dans l'adaptation des technologies de spatialisation sonore au contexte nomade. Outre la contrainte de taille, il faut aussi prendre en compte deux aspects supplémentaires : d'une part, la diversité des dispositifs et technologies disponibles sur le marché et d'autre part, la diversité des usages. Ces contraintes supposent, d'une part, que l'outil développé soit générique ou spécifique à un terminal et d'autre part, elles imposent une robustesse des techniques mises en œuvre pour garantir leur fonctionnement en dépit de l'hétérogénéité des environnements acoustiques, ainsi que des utilisateurs potentiels. Dans les solutions développées dans le cadre de ce travail de thèse, nous nous sommes attachés à satisfaire l'ensemble de ces points.

Le mémoire de thèse s'articule en six chapitres. Le premier chapitre présente un panorama des technologies de spatialisation sonore. Les outils de spatialisation sonore se servent des mécanismes perceptifs pour donner l'illusion à l'auditeur qu'il est immergé dans un espace où des sources sonores sont présentes. Avant de comprendre les techniques de spatialisation sonore, il est donc nécessaire de comprendre les mécanismes mis en jeu dans la perception de ce type de stimulus. Suite à cette première étape, une liste des différentes technologies de spatialisation disponibles aujourd'hui est dressée. Elle permet d'illustrer les différentes façons de mettre en œuvre la spatialisation sonore, soit en utilisant soit des approches physiques soit des mécanismes perceptifs.

Le second chapitre pose la question de l'intégration des technologies audio 3D dans les terminaux mobiles en évaluant l'ensemble des contraintes spécifiques à ce type d'équipement. Ce diagnostic est croisé avec l'inventaire des méthodes de spatialisation sonore afin de définir les techniques les plus à même d'être utilisées pour mettre en œuvre la spatialisation sonore dans les terminaux mobiles.

À l'issue de cet état des lieux et afin d'illustrer l'ampleur du problème à résoudre, une première expérience grandeur nature à l'opéra de Rennes (dans le cadre de l'événement "L'opéra dans tous les écrans") a été réalisée. L'un des objectifs était de tester une première maquette de dispositif nomade de captation et de restitution d'une scène sonore spatiale. Cette étude est décrite dans le chapitre III. Lors de cet événement, nous avons constitué une base de données d'enregistrements sonores utilisant des dispositifs variés de captation, tels que le binaural, l'ambisonique et la stéréophonie. Ces données nous ont servi à constituer un corpus d'échantillons sonores nous permettant de comparer en termes de perception différentes techniques de captation d'une scène sonore spatialisée en vue d'une restitution au casque. Les résultats de cette expérience ont été déterminants pour les choix du cap à suivre dans la suite de la thèse, car ils ont montré que les quatre capteurs du microphone *soundfield* permettent déjà une bonne captation et restitution de la scène sonore spatialisée, à condition d'effectuer un décodage adapté. Ils ont également

mis en évidence la faible qualité du rendu binaural des approches de décodage classique. En complément, une caractérisation objective a permis d'évaluer le décodage classique de l'ambisonique nous faisant opter par une approche de décodage actif.

Les conclusions de cette première étape ont montré que la prise de son spatialisée en contexte nomade est le point le plus critique. Nous nous sommes donc tout d'abord penchés sur cette question. Dans le chapitre IV, une première solution de prise de son 3D nomade est proposée et validée.

L'un des principaux objectifs de ce travail est de réduire au maximum la taille du dispositif de captation, ainsi que le nombre de capteurs. Utilisant des approches hybrides (technologie de spatialisation différente entre la captation et la restitution), la méthode développée tire parti des avantages de plusieurs méthodes pour obtenir des dispositifs relativement compacts et légers, compatibles avec le contexte de mobilité. Dans le but d'aller plus loin dans la démarche de simplification du dispositif de captation, nous avons souhaité fixer une limite maximale de trois capteurs pour la captation d'une scène sonore en 3D. Ce défi a été relevé avec un dispositif microphonique assez compact composé de trois capsules microphoniques cardioïdes. En complément du système microphonique, un post-traitement spatial est mis en œuvre afin d'identifier la direction des sources. Dans une première phase, la direction des sources est déterminée grâce à la directivité des capteurs et un signal sonore représentatif de la scène spatiale est extrait. Les informations extraites des capteurs permettent ainsi de caractériser complètement la scène sonore 3D et constituent une sorte de représentation au format "objet" de la scène. L'avantage de cette représentation "objet" est qu'elle est indépendante du dispositif de diffusion. En effet, grâce à un décodage adapté et spécifique au format de restitution, une image de la scène sonore ainsi captée peut être diffusée sur tout système de reproduction sonore spatialisée. Cette méthode permet ainsi une représentation efficace de l'information audio associée à une flexibilité technologique selon la méthode de restitution souhaitée. Pour aboutir à ces résultats, la meilleure configuration microphonique a tout d'abord été recherchée et une première méthode d'extraction des directions des sources a été déterminée. Des simulations numériques ont validé la méthode de localisation et le dispositif associé.

Compte tenu que cette première méthode implique des caractéristiques microphoniques assez précises, nous avons voulu proposer une nouvelle méthode permettant de prendre en compte les caractéristiques réelles non idéales des microphones, incluant leur potentielle interaction avec les équipements mobiles grand public. Cette seconde approche est présentée dans le chapitre V.

Le chapitre VI ouvre les perspectives des travaux de thèse. Afin de compléter la chaîne électroacoustique, des solutions de restitution sont notamment décrites afin de tirer parti des possibilités proposées par le concentré de technologie que représentent aujourd'hui les terminaux mobiles.

I Technologies audio 3D

I.1 Préambule

Ce chapitre a pour vocation de faire un état de l'art général des méthodologies utilisées dans la spatialisation sonore et de proposer au lecteur la "boîte à outils" comportant l'ensemble des techniques et des termes nécessaires à la compréhension des travaux abordés dans la suite du document. Les sections s'articulent de la façon suivante :

- une description des mécanismes mis en œuvre dans la perception des sources sonores dans l'espace est effectuée en I.2,
- un bref descriptif de la conception technique et de l'évolution des méthodes de spatialisation sonore replacées dans leur contexte historique depuis le XIX^{e} siècle jusqu'à nos jours est détaillée dans la première partie de I.3
- les méthodes énoncées dans la deuxième partie de I.3 sont ensuite décrites en détail à l'aide sous une classification en trois familles :
  - perceptives,
  - mixtes,
  - physiques.
- les méthodes décrites dans les premières sections, ayant atteint un niveau suffisant de maturité pour être utilisées dans le cadre de ce travail, ou ayant donné naissance à des dispositifs disponibles aujourd'hui sur le marché, sont abordées en I.5. Dans cette même section, seront abordés les points de convergence entre les différentes techniques et les formats de stockage et de transport. Ces formats sont considérés

comme des méthodes de représentation de la scène sonore spatialisée, et seront présentés au travers de trois familles :

- *channel based*,
- *soundfield based*,
- *object-based*,

la dernière de ces familles étant basée sur une description de la scène auditive par une localisation des sources.

- Les techniques usuelles d'analyse de la scène sonore seront abordées en A.
- enfin, en I.6 des méthodes permettant l'évaluation subjective de la qualité sonore seront abordées. En effet, les différentes technologies décrites dans ce chapitre, ayant pour objectif d'immerger un auditeur dans une scène sonore spatiale, il, est nécessaire de connaître les méthodes permettant de l'évaluer.

I.2 Rappels sur la localisation sonore

Parmi les cinq sens, l'audition est celui qui nous permet de percevoir les événements sonores. Généralement, les événements sonores se produisent dans l'espace et une localisation leur est attribuée. Il est alors possible de parler de perception sonore spatialisée [Blauert, 1983]. La localisation sonore est une expérience multisensorielle et dépend des connaissances et des attentes de l'auditeur. Sous certaines conditions, la position d'une source sonore est associée à la position d'un objet visuel si ce dernier est considéré comme producteur du stimulus. Ce phénomène, connu sous le nom d'"effet ventriloque" [Jack and Thurlow, 1973], est un exemple de la complexité du mécanisme de localisation sonore. Malgré le fait que la vue soit le sens privilégié pour la localisation des objets dans l'entourage, les événements sonores peuvent être perçus au-delà de notre champ de vision, ce qui donne à l'audition une position importante pour notre orientation, en particulier en cas d'alerte envers des agressions extérieures [Blauert, 2013].

En limitant la localisation sonore à une expérience monosensorielle, on considère uniquement les mécanismes quantifiables propres à l'audition. L'être humain est doté de deux capteurs de pression acoustique, les oreilles. Elles permettent à travers le système auditif de percevoir les perturbations de l'air dans une plage fréquentielle de 20 Hz à 20 kHz, avec une dynamique qui s'étend entre 0 dB et 120 dB [Fastl and Zwicker, 2007]. Schématiquement, ces deux capteurs se trouvent diamétralement opposés de part et d'autre d'un ellipsoïde et écartés d'environ 17 cm l'un de l'autre. Cette disposition conduit à des modifications du signal acoustique atteignant chaque oreille en fonction de

la position de la source. Les modifications du signal sont interprétées pour la localisation de la source comme des indices interauraux et spectraux. Les premiers interviennent notamment dans la localisation de la source en azimut et les seconds permettent de connaître précisément la position d'une source en élévation.

I.2.1 Indices Interauraux

Lorsque les deux oreilles sont mises à contribution dans la localisation sonore, on parle d'indices interauraux. En effet, une source placée en dehors du plan médian engendre des différences de temps et d'intensité entre les ondes atteignant les deux oreilles. Le mécanisme de localisation associé est inné et est prépondérant lorsque la source se trouve dans le plan horizontal. Ce mécanisme a été décrit par Lord Rayleigh comme *la théorie duplex*, où ces indices agissent conjointement dans la perception de la position latérale d'un son autour d'un auditeur [Rayleigh, 1907]. Ces indices sont de deux types : le premier détermine la différence interaurale de temps ou *Interaural Time Difference* (ITD) et le second détermine la différence interaurale d'intensité *Interaural Level Difference* (ILD).

I.2.1.a Différence interaurale de temps ITD

La différence de marche du trajet parcouru par une onde acoustique entre les deux oreilles engendre une différence de temps correspondant à l'ITD. L'ITD agit de deux manières distinctes en fonction de la plage fréquentielle. En effet, lorsqu'il s'agit d'une onde sinusoïdale continue, on parle alors de différence interaurale de phase ou *Interaural Phase Difference* (IPD) [Kuhn, 1977], mécanisme qui est utilisé pour les fréquences inférieures à environ 500 Hz à 600 Hz [Roth et al., 1980] [Dynes and Delgutte, 1992] [Zwislocki and Feldman, 2005]. Au-delà, c'est la différence de retard de groupe qui est prise en compte pour cette analyse. Le modèle sphérique de Woodworth [Woodworth et al., 1971] détermine assez précisément cet indice sur le plan horizontal par la relation

$$ITD_{HF} = \frac{a}{c}(\sin\theta + \theta), \tag{I.1}$$

où a est le rayon de la tête, c la célérité de l'onde acoustique et θ représentant l'azimut de la source par rapport au plan médian. D'autres modèles cherchant à représenter plus précisément la géométrie de la tête ont été proposés [Duda et al., 1999] [Nam et al., 2008] [Minnaar et al., 2000].

I.2.1.b Différence interaurale d'intensité ILD

La position des oreilles sur la tête définit un profil de directivité déterminé par l'obstacle qui représente la tête [Middlebrooks and Green, 1991]. L'ILD est l'indice décrivant la différence d'intensité entre les deux oreilles en fonction de la position de la source. Pour les fréquences dont la longueur d'onde est inférieure au diamètre de la tête, des phénomènes complexes de réflexion, diffusion et diffraction apparaissent en fonction de la position de la source et de la fréquence [Blauert, 1983]. Pour une source placée sur l'axe interaural, l'onde incidente est complètement réfléchie, engendrant un gain de +6 dB pour l'oreille ipsilatérale [Guillon, 2009], c'est-à-dire se trouvant directement sur l'axe de la source. Ce phénomène se produit lorsque la longueur d'onde est inférieure à la distance interaurale. De plus, les pavillons modifient le niveau sonore en fonction de l'angle d'incidence de la source et de sa fréquence. Plusieurs auteurs suggèrent que l'ILD exploite ces différences spectrales *Interaural Spectral Difference* (ISD) pour lever les ambiguïtés produites par l'utilisation de l'ITD seule [Larcher, 2001]. Néanmoins, des expériences psychoacoustiques suggèrent que la variation spectrale est utilisée de façon monaurale par l'oreille ipsilatérale, du moins pour les sources latéralisées. On parle alors d'indices monoauraux [Guillon, 2009].

I.2.2 Indices spectraux et HRTF

Les **indices spectraux** ou **indices monauraux** sont considérés comme tels, car, d'une part, il a été prouvé qu'ils agissent sur la localisation des sources de façon monaurale, notamment pour la détermination de la position en élévation d'une source. D'autre part, la condition d'écoute binaurale n'apporte pas de précision pour la localisation dans cette coordonnée [Middlebrooks and Green, 1991]. Néanmoins, certaines études ont formulé l'existence d'une ILD spectrale telle que décrite par Guillon [Guillon, 2009].

Les réflexions et diffractions de l'onde acoustique sur le torse, la tête et notamment les pavillons des oreilles, modifient le signal sonore atteignant chaque tympan. Ces modifications peuvent être interprétées comme un filtrage du signal engendré par la source.

Si nous considérons une source S qui émet un signal temporel $s(t)$, les signaux perçus à l'oreille gauche et droite de l'auditeur sont respectivement $s_l(t)$ et $s_r(t)$ où

$$s_{l,r}(t) = s(t) * h_{l,r}(t). \tag{I.2}$$

Dans cette relation $h_{l,r}(t)$ définit la réponse impulsionnelle correspondant à la direction de la source. Dans un environnement anéchoïque, la morphologie de la tête (incluant le torse et les oreilles) est la seule responsable de $h_{l,r}(t)$. On parle alors de réponse impulsionnelle liée à la tête *Head Related Impulse Response* (HRIR) (figure I.1). Dans le

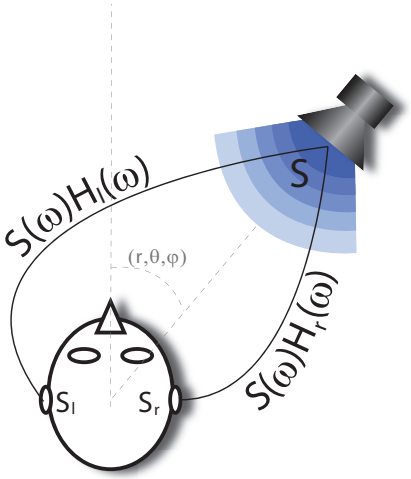


Figure I.1 : Signaux perçus à l'entrée des oreilles après avoir été modifiés par leur trajet et la morphologie de la tête.

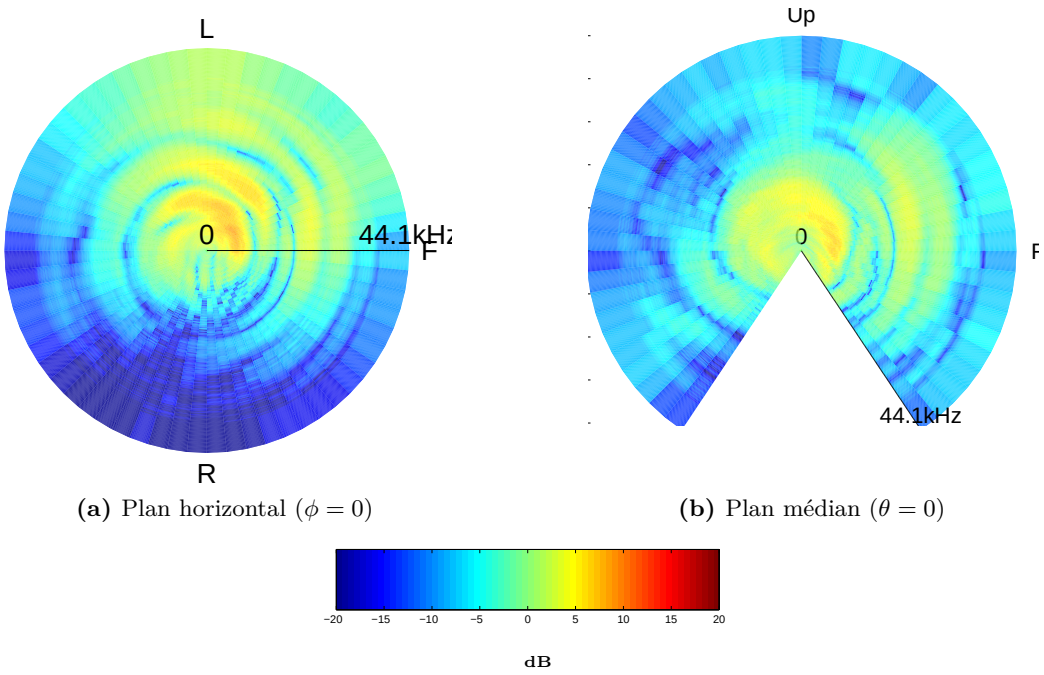
domaine fréquentiel, l'équation I.2 devient

$$S_{l,r}(\omega) = S(\omega)H_{l,r}(\omega), \tag{I.3}$$

où $H_{l,r}(\omega)$ est la fonction de transfert en fréquence liée à la tête ou *Head Related Transfert Function* (HRTF).

Les fonctions $h_{l,r}(t)$ et $H_{l,r}(\omega)$ varient en fonction de la direction de la source et correspondent au filtrage directionnel créé par la tête, et notamment les pavillons des oreilles. Ce phénomène engendre des colorations du signal, permettant la localisation des sources en élévation. La figure I.2 affiche une représentation du spectre d'amplitude des HRTF dans le plan horizontal et dans le plan médian. Dans une approche plus large, on peut considérer un jeu de HRTF comme la base de données comportant toute l'information nécessaire pour la localisation d'une source, car une paire de HRTF possède également les informations correspondant aux indices interauraux définis précédemment. Chaque jeu de HRTF dépend directement de la morphologie de l'auditeur et définit l'empreinte auditive propre à chacun.

Cependant, la morphologie changeant tout au long de la vie, un réapprentissage permanent de ces indices est indispensable. D'autre part, la plasticité du cerveau permet également l'apprentissage de nouveaux jeux de HRTF engendrés par des changements morphologiques radicaux [Van Wanrooij, 2005, Hofman et al., 1998]. Dans son analyse directionnelle, le système nerveux exploite les informations prépondérantes dans le spectre telles que les pics et les creux sur des plages de fréquences plus ou moins larges. Même si



(a) Plan horizontal ($\phi = 0$) **(b)** Plan médian ($\theta = 0$)

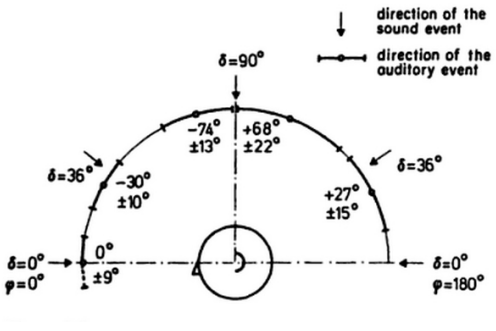
Figure I.2 : Représentation spatiale de l'amplitude du spectre des HRTF sur le plan horizontal (a) et le plan médian (b). Les basses fréquences sont représentées au centre et les hautes fréquences en périphérie selon une échelle de couleurs en dB. HRTF de l'oreille gauche du sujet n°1 de la base de JM Pernaux [Pernaux, 2003].

aujourd'hui il est reconnu que les accidents spectraux situés dans les hautes fréquences (engendrés par les pavillons des oreilles) sont les principaux responsables de la localisation en élévation [Langendijk and Bronkhorst, 2002], il a également été démontré que l'absence d'énergie dans les basses fréquences dégrade la localisation tant en condition binaurale que monoaurale [Butler and Humanski, 1992].

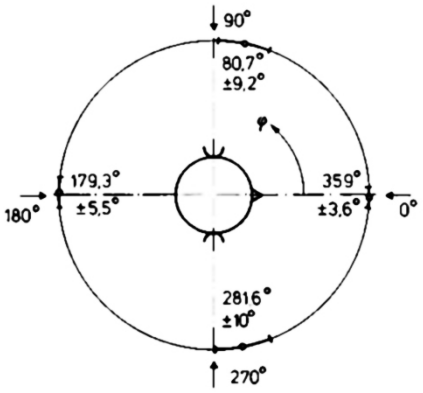
I.2.3 Indices de location dynamique

Lorsqu'on cherche à localiser une source sonore émettant un signal continu, la tête est tournée de façon involontaire en direction de la source pour accomplir la tâche [Thurlow and Runge, 1967] scrutant l'espace afin de trouver la position qui minimise les différences interaurales [Middlebrooks and Green, 1991]. En effet, les deux indices binauraux varient de façon opposée avec les rotations de la tête autour de l'état d'équilibre quand la source se trouve sur le plan médian.

Lorsque le stimuli est bref, il devient difficile pour l'auditeur de faire face à la source, mais il a été démontré que les mouvements de la tête permettent une meilleure localisation en azimut pour des événements sonores de courte durée [Perrett and Noble, 1997]. En effet,



(a) Localisation et flou de localisation dans le plan médian d'une voix continue familière d'après l'étude de Damaske et Wagener [Damaske and Wagener, 1969]



(b) Localisation et flou de localisation dans le plan horizontal pour des salves de bruits blancs de 100 ms d'après l'étude de Preibich-Effemberg [Preibisch-Effenberger, 1966]

Figure I.3 : Représentation du flou de localisation dans le plan horizontal et le plan médian [Blauert, 1983].

les indices interauraux définissent des ensembles d'isovaleurs définies géométriquement sur ce qu'on appelle des cônes des confusion [Wallach, 1939]. Les indices spectraux permettent de lever en grande partie ces ambiguïtés, mais il reste parfois des ambiguïtés hémisphériques avant - arrière (sur plan frontal passant par la ligne définie par les deux oreilles), qui sont levées grâce aux mouvements relatifs de la tête par rapport à la source comme l'a démontré Wightman [Wightman and Kistler, 1999], confortant l'hypothèse de Wallach [Wallach, 1940]. Des rotations de 5° à une vitesse de $50°s^{-1}$ permettent d'extraire des indices de localisation suffisants pour lever ces ambiguïtés [Macpherson, 2009], même pour des sons de courte durée.

Aujourd'hui, il reste difficile de connaître le poids joué par les indices de localisation dynamique dans la localisation des sources. Blauert [Blauert, 1983] les considère comme secondaires dans une notion de hiérarchie entre les différents indices, plaçant les indices spectraux au premier rang. En effet, les expériences de localisation menées par Faure montrent que les indices dynamiques n'améliorent pas la localisation avec des indices spectraux modifiés [Faure, 2005].

I.2.4 Précision dans la localisation sonore

La précision dans la localisation sonore peut être étudiée de deux manières. D'une part, elle est déterminée comme la distance entre la position de la source et celle du percept sonore qui en résulte. D'autre part, elle est évaluée comme la distance permettant de discriminer deux événements sonores distincts. Ces deux cas ont été appelés par Blauert

Référence	Nature du signal	type de test	Flou de localisation (°)
Klemm (1920)	Impulsionnel(Clics)	-	0.75 - 2.00
King et Laird (1930)	Impulsionnel(Clics)	I	1.6
Steves et Newman (1936)	monochromatique	D	4.4
Schmidt et al (1953)	monochromatique	-	>1
Sandel et al. (1955)	monochromatique	I	1.1 - 4.0
Mills (1958)	monochromatique	D	1.0 - 3.1
Stiller (1960)	Bruit à bande étroite, salves de sinus	D	1.4-2.8
Boerger (1965)	Bruit Gaussian, salves de sinus	I	0.8 - 3.3
Gardner (1968)	Parole	I	0.9
Perrot (1969)	Salves de sinus avec enveloppe temporelle	D	1.8-11.8
Blauert	Parole	D	1.5
Haustein et Schirmer (1970)	Bruit large bande	D	3.2

Tableau I.1 : Recueil des mesures du flou de localisation d'après [Blauert, 1983]. Le type de test est identifié par I pour identification ou D pour discrimination.

[Blauert, 1983] flou de localisation (*localization blur*) (figure I.3). Les performances de localisation ont été largement étudiées grâces à des tâches d'identification, lors desquelles il est demandé à l'auditeur de localiser une source de manière absolue dans l'espace [Oldfield and Parker, 1984, Makous and Middlebrooks, 1990, Carlile et al., 1997] et par des tâches de discrimination. Les méthodes d'identification permettent d'obtenir des valeurs pertinentes du point de vue statistique [Letowski and Letowski, 2011], mais ne permettent pas d'estimer le seuil de discrimination entre deux sources distinctes. Grantham [Grantham, 1995] signale que la méthode la plus directe pour mesurer le flou de localisation est celle du paradigme expérimental de discrimination du *Minimum Audible Angle* (MAA) de Mills [Mills, 1958], donnant ainsi une valeur du seuil d'acuité auditive. La valeur obtenue est dépendante du type de stimulus, de son contenu spectral, ainsi que de sa direction (tableau I.1 [Blauert, 1983]). Des expériences plus récentes [Daniel, 2011] montrent que le flou de localisation dépend également du niveau sonore et de la présence d'une source distractrice. Généralement, elle est considérée de l'ordre de 1° à 2° en position frontale et atteint des valeurs entre 8° et 10° [Kuhn, 1987] en position latérale pour décroître encore à l'arrière, atteignant des valeurs de 6° à 7° [Mills, 1958, Perrott, 1969]. Dans les basses fréquences, dans une position frontale, la valeur du MAA correspond au seuil de l'ITD (*Just-noticeable Difference* (JND))

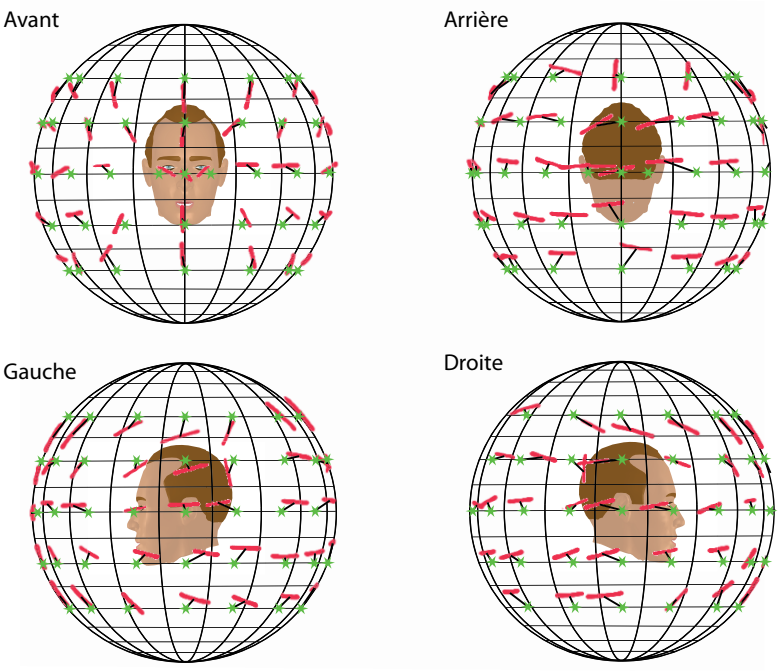


Figure I.4 : Illustration des performances de localisation (d'après [Carlile et al., 1997]). Les étoiles vertes matérialisent les directions des sources à localiser, tandis que les longueurs des traits rouges expriment la dispersion des résultats dans la direction d'étirement maximal (extrait de [Guillon, 2009].).

($\simeq$ 10 μs) et dans les hautes fréquences, il correspond au seuil (JND) de ILD (0,5 - 1.0 dB), pour atteindre des valeurs supérieures à des fréquences moyennes [Mills, 1958]. Le MAA pour la localisation en élévation oscille entre 3° et 9° pour une position frontale [Perrott and Saberi, 1990, Blauert, 1983] (figure I.3a).

Lorsqu'une même analyse est effectuée avec une tâche d'identification absolue, les valeurs du flou de localisation peuvent être représentées par l'écart type de la distribution des réponses de la tâche. Dans ce cas, selon Carlile [Carlile et al., 1997], la valeur est de l'ordre de 5° en position frontale et atteint des valeurs de 10° à l'arrière. En élévation, la dispersion est plus forte pour les élévations de ±40°. Le flou de localisation en élévation est compris entre 10° et 12° quel que soit l'azimut (figure I.4).

I.2.5 Localisation de plusieurs sources

Ce qui a été exposé dans les paragraphes précédents vaut pour la localisation d'une source en champ libre, même si, en réalité, il faut prendre en compte l'influence de l'environnement acoustique.

En présence d'obstacles, les réflexions sur ces derniers peuvent en effet être modélisées

comme des sources secondaires agissant directement sur le champ sonore. D'un point de vue perceptif, ces sources secondaires deviennent concurrentes de la source principale, modifiant ainsi les événements sonores perçus. En effet, le signal émis par la source secondaire est cohérent (du point de vue du signal) avec le signal secondaire. Lorsque ces deux signaux sont perçus par un auditeur, l'événement acoustique n'est plus associé à la position de la source sonore qui le génère.

Blauert [Blauert, 1983] dénombre trois cas perceptifs distincts dans le cas de deux sources cohérentes :

- un seul événement sonore est perçu à une position déterminée par la position et l'intensité des sources sonores,
- un seul événement sonore est perçu à une position déterminée par la position d'une seule des sources,
- deux événements sonores distincts sont perçus.

Le premier cas est obtenu lorsque le retard entre les deux signaux émis est faible. Un événement sonore est perçu à une position intermédiaire entre les sources. Sa position est dépendante du retard ainsi que de l'intensité des signaux. Ce phénomène est connu sous le nom de "sources fantômes" (ou "phantom sources" en anglais) et est le principe de base de la stéréophonie.

Le deuxième cas se produit lorsque le retard est supérieur à 1 ms à la position de l'auditeur et l'événement sonore est perçu à l'emplacement de la première source. Ce phénomène est connu sous le nom d'effet du premier front d'onde, effet de précédence ou effet Haas. L'événement sonore résultant est perçu principalement avec des modifications timbrales [Grantham, 1995].

Le troisième cas apparaît lorsque le retard devient très important ($>$ 30 - 35 ms). Dans ce cas, chaque événement sonore est perçu de façon indépendante et la source secondaire est perçue comme un écho de la première, chaque source étant localisée à son propre emplacement.

Le retard déterminant l'apparition de deux sources indépendantes a été appelé par Blauert [Blauert, 1983] le seuil d'écho. La valeur de ce seuil est fortement dépendante du stimulus et de son enveloppe temporelle. Par exemple, il atteint des valeurs de 35 ms pour la parole et de 5 ms pour des signaux impulsionnels.

Les performances de localisation ainsi que le MAA sont dégradées d'environ 10 à 15 % avec l'utilisation de signaux monochromatiques, car l'effet de précédence ne peut pas être actif [Grantham, 1995] dans ces conditions. Rakerd et Hartmann le confirment dans une tâche de localisation absolue de signaux continus et de signaux impulsionnels

[Rakerd and Hartmann, 1992]. L'expérience est effectuée tout d'abord dans un environnement anéchoïque et ensuite, en présence de surfaces réfléchissantes. Il a été montré, d'une part, que les signaux continus sont mieux localisés dans l'environnement anéchoïque, avec une erreur de 3.5° contre 9.7° en présence de surfaces réfléchissantes. D'autre part, la localisation des signaux impulsionnels présente un écart-type de seulement 3.0° et 4.4° respectivement pour chaque condition, démontrant ainsi l'importance du phénomène de précédence pour la localisation des signaux impulsionnels uniquement.

I.2.6 Perception de la distance

Par rapport à l'attention portée à la perception de la direction des sources, la perception de la distance a été peu étudiée. Les mécanismes cognitifs permettant d'extraire cette information sont basés sur un ensemble d'indices acoustiques disponibles et dépendent fortement de la familiarité des signaux utilisés ainsi que des informations acquises par apprentissage [Grantham, 1995].

Blauert [Blauert, 1983] dénombre quatre indices principaux permettant la perception de la distance :

- le niveau sonore,
- le champ réverbéré par rapport au champ direct,
- la modification spectrale,
- les indices interauraux.

En champ libre, l'atténuation géométrique de 6 dB par doublement de la distance donne un indice fiable en présence d'une référence connue par l'auditeur. Le seuil perceptif à partir de 6 m correspond à une variation du 5% de la distance, ce qui est équivalent au seuil de variation de la sonie de 0.5 dB. Pour les distances inférieures à 6 m, le seuil augmente à 20% de la distance relative, valeur qui ne correspond plus à la sonie [Strybel and Perrott, 1984]. Dans un environnement réverbéré, la décroissance géométrique n'est plus respectée et l'atténuation est plus faible. Dans ce second cas, c'est l'énergie relative du champ direct par rapport au champ réverbéré qui apporte l'information nécessaire pour la localisation de la distance. Lorsque les distances deviennent importantes, le troisième indice apporte une information supplémentaire par rapport à l'éloignement de la source. Il est important de souligner que l'atténuation de l'air dans les hautes fréquences est relativement faible (de l'ordre de quelques dB par 100 m) [Zahorik, 2002]. La modification spectrale est aussi rencontrée dans des environnements réverbérants, où les premières réflexions apportent des interférences sur le signal perçu, modifiant principalement son timbre.

Les différences interaurales ne sont pas dépendantes de la distance lorsque la source est suffisamment éloignée de l'auditeur (hypothèse d'ondes planes). Cependant, lorsque la source s'approche de l'auditeur, cette hypothèse n'est plus valable et les différences interaurales dépendent également de la distance donnant ainsi des indices supplémentaires pour sa perception.

I.3 Tour d'horizon des technologies de spatialisation sonore

Dans la section précédente, les stratégies utilisées par le système auditif pour localiser des sources sonores ont été décrites. La localisation des événements sonores étant purement perceptive, il est possible de créer virtuellement des informations sonores permettant la création de sources virtuelles et de les disposer ainsi dans l'espace.

Un espace tridimensionnel peut être caractérisé en termes de largeur, hauteur et profondeur. La spatialisation sonore a comme objectif de créer des signaux sonores au niveau des oreilles de l'auditeur en lui donnant l'illusion de sources sonores repérables dans ces 3 dimensions.

Grâce au traitement du signal et à la compréhension des phénomènes acoustiques impliqués dans la localisation acoustique (I.2), il est possible de modeler le son afin de s'approcher des mécanismes nous permettant de localiser une source sonore de manière naturelle [Blauert, 1983].

I.3.1 Des premiers pas au son multicanal 5.1

I.3.1.a Prémices

La première expérience permettant une restitution sonore contenant une information spatiale remonte à la fin du XIX^{e} siècle et a été réalisée par Clément Ader. En 1881, pendant l'Exposition de l'Electricité, cet ingénieur, pionnier de l'aviation et créateur du premier réseau téléphonique parisien, a installé 80 postes téléphoniques devant la scène de l'Opéra Garnier. Le son ainsi capté était restitué de façon binaurale au Palais des Expositions. Dès 1890, ce système a été utilisé de façon commerciale permettant aux abonnés d'écouter l'opéra depuis leur maison (figure I.5) [Du Moncel, 1887, Lange, 2014, Laster, 1983, Vaslin, 2010, Ltd, 2014].

La spatialisation sonore a trouvé ensuite une application pratique lors du premier conflit mondial [museum of retro technology, 2009] [Mechanics, 1938] avec le développement d'appareils acoustiques permettant la localisation des avions ou des tranchées ennemies (figure I.6).

Ce n'est pour autant qu'en 1925 que Kapeller effectue la première diffusion stéréophonique hertzienne à l'opéra de Berlin sur deux fréquences AM en simultané. Aux États-Unis, Doolittle produit la même expérience la même année sur les ondes de la WPAJ. Doolittle

(a) Affiche publicitaire du service du théâtrophone, *La parisienne au théâtrophone*, Jules Chéret (1896).

(b) Théâtrophone à pièces, Gravure Victorienne. XIX^{e}

Figure I.5 : Illustrations du théâtrophone, d'après [Lange, 2002].

précise à ce moment que l'utilisation d'un casque est essentielle pour retrouver la "naturalité" des enregistrements [Sunier, 1960].

En 1930, Blumlein développe pour EMI le premier procédé d'enregistrement de deux pistes sur un disque à microsillon [Vanderlyn, 1978] et en 1954, Livingston Audio commercialise les premiers enregistreurs multipistes sur bande magnétique de 1/4 de pouce. La stéréophonie devient le standard de diffusion sonore jusqu'à nos jours malgré les avancées de la restitution multi haut-parleurs exploitant à la fois des mécanismes perceptifs et des formalismes physiques [Sunier, 1960].

I.3.1.b Stéréophonie

Du grec ancien "stereos", qui signifie solide, et, par extension, volume ou espace, et, "phone" qui signifie voix ou son, le mot "stéréophonie", comme sa définition étymologique l'indique, a été donné à la représentation du son dans l'espace. Par usage, ce terme est devenu l'appellation de l'enregistrement et la restitution sonore sur deux haut-parleurs.

Les premiers jalons de cette technique ont été posés aux Bell Labs à la fin du XIX^{e} siècle par Snow et Steinberg, qui définissent le système de prise stéréophonique idéal

Figure I.6 : Radar acoustique bicône à Bolling Field, USA, 1921, d'après [museum of retro technology, 2009].

comme un rideau de microphones placés devant la scène sonore. Plus tard, Moire et Olson, démontrent qu'une restitution idéale de la scène sonore ainsi captée doit être faite par un réseau de haut-parleurs disposés de façon à respecter la distribution spatiale du champ acoustique capté [Moir, 1958] [Olson, 1978]. A cause des contraintes techniques de l'époque, Snow et Steinberg ont dû limiter le nombre de transducteurs à trois. Les prix des amplificateurs et de l'électronique en générale ont fait que ce nombre de haut-parleurs soit limitant pour une application grand public. C'est pour cela qu'au milieu du XX^{e} siècle, lorsque l'utilisation de deux haut-parleurs est devenue une réalité, ce format a été adopté à l'unanimité comme le standard en diffusion sonore, et le nom de stéréophonie a été attribué à toute restitution sonore sur deux haut-parleurs au lieu du terme duophonie ou biphonie, qui aurait été plus exact.

Le son stéréophonique utilise deux haut-parleurs placés face à l'auditeur et exploite grossièrement les indices binauraux (I.2.1), générant des différences de temps et/ou d'amplitude (indépendamment ou simultanément) entre les deux signaux, afin de créer l'image d'une source sonore dans l'espace délimité par les deux haut-parleurs, source dite "source fantôme".

Afin de créer cet effet, plusieurs techniques de captation ont été mises au point. La capta-

tion peut être réalisée à l'aide de deux microphones positionnés de façon coïncidente ou écartés d'une certaine distance. Lorsque les microphones sont coïncidents, leur directivité est exploitée pour obtenir une différence de niveau entre les canaux droit et gauche. Autrement, lorsqu'ils sont écartés, c'est la différence de temps d'arrivée du signal sur les deux capsules qui assure le positionnement de la source fantôme (I.2.5).

Le premier cas est exploité par la technique stéréophonique A-B utilisant deux microphones omnidirectionnels écartés d'environ 0,5 m dans le but d'engendrer un retard entre 0 et 1,5 ms en fonction de l'angle de la source (figure I.7a).

Le deuxième cas se produit dans la configuration X-Y. Cette méthode utilise deux microphones cardioïdes coïncidents et écartés de 45° (figure I.7b) ou, dans le cas de la méthode appelée Blumlein, deux microphones bidirectifs (figure I.7c). Ces techniques ont été développées pour générer des différences d'intensité entre les deux signaux stéréphoniques en fonction de l'angle de provenance de la source, et dans un souci de compatibilité avec un rendu monophonique. Il est à remarquer que les deux canaux étant coïncidents, leur addition n'engendre pas d'interférences destructives gênantes pour le "downmix [1]" monophonique.

D'autres méthodes comme la MS stéréo[2] utilisant un microphone omnidirectionnel et un microphone bidirectif ont été proposées (figure I.7d). Cette méthode a pour objectif de générer par matriçage les signaux équivalents à ceux issus d'un couple cardioïde coïncident et pointant dans des directions opposées (figure I.7e).

Des méthodes mixtes comme celle du couple ORTF cherchent à la fois à introduire un retard entre les signaux et une différence d'amplitude grâce à la directivité des capteurs (figure I.7f). Ce couple est composé de deux microphones cardioïdes écartés de 0,17 m avec un angle d'ouverture de 110° afin d'optimiser [Farina and Tronchin, 2005] les indices interauraux de localisation, à la fois temporels et d'amplitude (I.2.1).

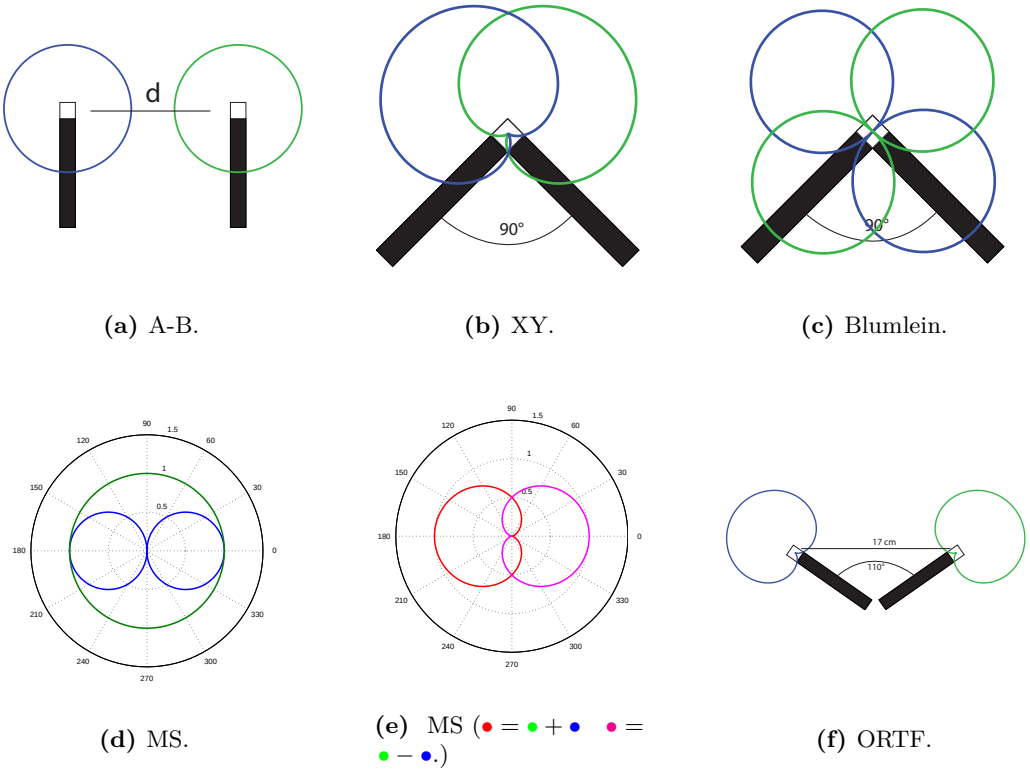
Il est aussi possible de créer un mixage stéréo à l'aide de plusieurs signaux monophoniques en utilisant le procédé de "pan pot" ou panoramique d'intensité. Le procédé consiste à augmenter l'intensité sonore sur l'une des deux voies pour déplacer l'image sonore entre les deux haut-parleurs.

La principale limitation de la technique stéréophonique est la faible largeur de la zone d'écoute ou *sweetspot*. L'effet de précédence et notamment l'effet Haas, ou loi du premier front d'onde (I.2.5), fait que lorsque l'auditeur n'est pas placé à une distance identique des deux haut-parleurs, l'ITD augmente et la source est perçue à la position du haut-parleur le plus proche, détruisant alors l'effet de la source fantôme [Litovsky and Macmillan, 1994].

La première restitution stéréophonique grand public remonte à 1939, réalisée lors de

[1] Réduction de canaux

[2] Aussi connu comme *Mid-Side stereo* (US) *Middle-Side stereo* (UK) ou *Mitte-Scite stereo* (D).



(a) A-B. **(b)** XY. **(c)** Blumlein.

(d) MS. **(e)** MS (● = ● + ● ● = ● − ●.) **(f)** ORTF.

Figure I.7 : Configurations microphoniques des techniques de captation stéréophonique.

la foire de New York. A cette occasion, les ingénieurs des Bell Labs intègrent une restitution stéréophonique au court métrage stéréoscopique de John Norling *In Tune with Tomorrow*, permettant ainsi de recueillir les impressions des cinq millions de spectateurs [Sunier, 1960, Zone, 2012].

Au même moment, Garity et Hawkings étudient comment réaliser la bande-son du film *Fantasia* de Walt Disney qui sortira en 1940. Walt Disney, soucieux de la qualité sonore du film musical *Fantasia*, et insatisfait de la qualité des supports audio de l'époque, développe le procédé appelé *Fantasound*. Afin d'exploiter au mieux la dynamique des supports optiques existants, les ingénieurs de Disney optent pour des enregistrements sur plusieurs canaux optiques. Lors de la reproduction, le niveau du gain des différents amplificateurs est piloté par une piste de contrôle permettant ainsi d'accroître la dynamique des enregistrements. Finalement, les contraintes d'exploitation ont fait que le film n'a été principalement diffusé qu'en son monophonique. Ce n'est que dans les années 1950, lorsque la stéréo devient le standard de la diffusion cinématographique, que le son stéréophonique du film est restauré à partir des bandes originales [Smith, 2007]. A la même époque, l'apparition du disque stéréo à microsillon, de la société Audio Fidelity Records en 1957, place le son stéréophonique comme le nouveau standard de la diffusion musicale, et ce jusqu'à nos jours.

A la radio, la première expérience de stéréophonie a été réalisée par la BBC en décembre 1925 avec la transmission sur deux fréquences du même programme, afin que leurs auditeurs puissent écouter à l'aide de deux postes les signaux gauche et droite sur les fréquences respectives. Pendant la deuxième moitié de la décennie de 1950, les premiers tests de FM stéréo débutent aux Etats-Unis et ce n'est qu'en août 1962 que la BBC réalise sa première transmission en FM stéréo.

La télévision s'est également servie de la radio pour la diffusion de l'audio en stéréo. A noter que ce n'est qu'en 1978 que la télévision japonaise introduit le multiplexage de l'audio et la vidéo [Pollak, 1984].

I.3.1.c Stéréophonie multicanal

A titre d'extension à la stéréo, d'autres dispositifs ont vu le jour en incrémentant le nombre de haut-parleurs, créant ainsi la stéréophonie multicanal. Le premier système sonore multicanal grand public a été proposé dans les années 1970. La quadraphonie ou le 4.0 utilise quatre canaux, deux frontaux et deux à l'arrière, les signaux envoyés à chaque haut-parleur étant plus ou moins indépendants en fonction du format et du décodage utilisé. Cette technologie a été principalement utilisée pour les enregistrements musicaux. Un système équivalent en nombre de canaux a été utilisé pour la diffusion cinématographique, le *Left, Center, Right, Surround* (LCRS). Ce système a la particularité d'utiliser trois haut-parleurs frontaux et un haut-parleur à l'arrière.

Dans le souci d'améliorer l'enveloppement sonore pour la diffusion cinématographique, deux canaux supplémentaires ont été ajoutés produisant ainsi ce qu'on connaît aujourd'hui sous le nom de système 5.1. Ce système comporte trois canaux frontaux, deux arrières et un canal dédié aux basses fréquences. En 1976, ce système a été utilisé pour la première fois en détournant l'utilisation des six canaux audio de la bande magnétique des films 70 mm. Jusqu'à nos jours, cette technologie a été largement utilisée pour la diffusion cinématographique en salle et par les dispositifs *home-cinema* grand public, et est devenue un standard ITU en 2004 avec la norme UIT-R BS 775-1 [IUT, 2012].

Par extension au système 5.1, des nouveaux haut-parleurs ont été ajoutés créant ainsi des systèmes comme le 7.1, le 10.2 ou le 22.2 [Hamasaki et al., 2005].

Les ingénieurs du son travaillant sur ce type de support utilisent majoritairement les deux canaux frontaux pour diffuser l'environnement sonore présent à l'écran et le canal central pour les dialogues. Les canaux placés à l'arrière et sur les côtés sont surtout exploités pour diffuser des bruits d'ambiance généralement diffus et non présents à l'écran. Avec les formats prenant en compte la diffusion en élévation, les créateurs de contenu sont maintenant confrontés au flou de localisation sur cette coordonnée (I.2.4) et utilisent majoritairement des sources sonores indépendantes et décorrélées des signaux diffusés dans le plan horizontal.

Parallèlement, de nombreux systèmes de diffusion multicanal ont été développés dans le cadre de la musique concrète comme "l'acousmonium" de François Bayle ou les expériences de Stockhausen, Varese ou Xenakis [Zanesi and Gayou, 1983, Prager, 2012, Direction générale des relations culturelles, 2011, Ircam-Centre Pompidou, 2009], [Ircam-Centre Pompidou, 2012].

Aujourd'hui, grâce à l'apparition de la Radio Numérique Terrestre, de la diffusion sonore sur Internet et du DVD audio, les réalisateurs de contenus sonores peuvent exploiter ces nouveaux formats comme supports de création grâce à des initiatives comme celle de "Nouvoson" de Radio France [radiofrance.fr, 2014].

I.3.2 Le son 3D en évolution

Au tournant du XX^{e} siècle, le domaine du son 3D connaît plusieurs évolutions majeures.

Tout d'abord, les avancées en termes de codage des signaux sonores, ainsi que les capacités de stockage et transmission des données, font du son spatialisé une réalité pouvant atteindre le grand public. Aujourd'hui, il est possible de trouver chez des particuliers des systèmes de diffusion sonore semblables à ceux équipant les salles de cinéma. La réponse des géants en la matière (MPEG, Dolby et DTS) ne s'est pas faite attendre. Afin d'attirer dans les salles un public friand d'expériences immersives, ils ont introduit un nombre croissant de canaux audio, grâce à des configurations comme le 10.2, le format Auro 3D et le Dolby Atmos. A chacune de ces configurations, un encodage et un décodage des signaux sonores sont associés. Pour éviter aux créateurs les mixages multiples, et afin de réduire les coûts de production, certains de ces encodages permettent une adaptation des signaux au système de restitution comportant plus ou moins de haut-parleurs que ceux prévus lors de la production initiale (*upmix* [3] et *downmix*).

On peut également noter l'émergence de technologies jusqu'aujourd'hui reléguées à des prototypes de laboratoire telles que l'holophonie et l'ambisonique.

Ces deux méthodes, basées sur la reconstruction physique de l'onde acoustique, sont à la fois similaires et différentes et peuvent être également complémentaires. D'une part, l'holophonie basée sur le principe de sources élémentaires de Huygens permet une restitution du champ sonore sur une zone élargie. De l'autre, l'ambisonique permet la captation en l'encodage spatial de ce même champ acoustique sur une base de vecteurs propres.

Toutes ces méthodes de restitution nécessitent un grand nombre de transducteurs et font de la technologie binaurale une alternative attractive, permettant de s'affranchir des

[3] Adapter des signaux sonores comportant un nombre signaux N à un dispositif ou une technique comportant un nombre N' de signaux, $N' > N$. Les informations manquantes sont extraites des pistes sonores existantes.

contraintes qu'elles imposent. En effet, utilisant uniquement deux canaux, cette technique permet la reproduction des signaux à l'entrée des conduits auditifs pour créer une image sonore 3D grâce à l'utilisation d'un "simple "casque stéréophonique.

I.4 Principales familles de technologies son 3D

Les méthodes de spatialisation sonore ou son 3D peuvent être classées en deux groupes. Le premier exploite les mécanismes de perception auditive afin de restituer des images sonores et le deuxième tente de reconstruire un champ acoustique autour de l'auditeur à l'aide de formalismes physiques et mathématiques.

Il est possible d'introduire une troisième catégorie combinant les deux approches. Dans cette catégorie, la méthode binaurale a toute sa place, car elle reconstruit le champ acoustique à proximité des oreilles de l'auditeur à l'aide d'un casque exploitant les indices de localisation sonore.

Dans les paragraphes de cette section, nous allons décrire ces technologies dans l'ordre suivant :

- méthodes perceptives,
- méthodes mixtes,
- méthodes physiques.

I.4.1 Méthodes perceptives

Cette famille inclut principalement les méthodes stéréophoniques et stéréophoniques multicanales qui ont été largement décrites en I.3.1.b.

Pour rappel, ces techniques exploitent les indices interauraux naturels de localisation sonore présentés en I.2 de manière simplifiée. Des sources fantômes sont générées à des points intermédiaires situés entre les haut-parleurs les diffusant par une restitution grossière des indices interauraux (I.2.1).

Figure I.8 : Premier enregistrement binaural effectué par Fletcher avec la tête acoustique "Oscar".

I.4.2 Une méthode mixte : la technologie binaurale

Cette technique peut être qualifiée de mixte car elle a pour but de recréer l'onde acoustique générée par une source uniquement à l'entrée des conduits auditifs. L'onde sonore ainsi constituée comporte l'ensemble des indices naturels utilisés dans la perception sonore spatialisée à la fois monauraux et interauraux (I.2).

I.4.2.a Prise de son binaurale

L'utilisation de la technologie binaurale remonte à l'hiver 1932, quand, pour la première fois, Fletcher réalise une prise de son à l'aide de deux microphones placés dans les oreilles d'un mannequin en bois « Oscar »[Fletcher, 1934] (cf figure I.8). Cette expérience fera l'objet d'une démonstration lors de l'Exposition Universelle de Chicago avec une retransmission d'une pièce de Stokowski. A cette occasion, la retransmission se fait par voie filaire dans une salle à Philadelphie. C'est suite à cette expérience, qu'on fait pour la première fois le distinguo entre prise de son stéréophonique multipiste ou dichotique et prise de son binaurale [Sunier, 1960].

Cette technique peut être implémentée dès la prise de son en utilisant deux méthodes. La première consiste à placer une paire de microphones au niveau de l'entrée du conduit auditif du preneur de son. Elle permet ainsi de capter en même temps que les signaux sonores, l'ensemble des indices naturels de localisation. La prise de son peut être entachée des mouvements de la tête et des bruits parasites tels que la respiration ou la déglutition de la personne équipée du dispositif. Afin de réduire ces problèmes, une deuxième méthode

(a) Kemar.

(b) HATS 4128-C de B&K.

(c) HMS IV de Head Acoustics.

(d) KU 100 de Neumann.

Figure I.9 : Têtes acoustiques.

peut être mise en œuvre, utilisant une tête artificielle équipée de microphones au niveau des oreilles. Plusieurs modèles simulant la morphologie humaine de façon plus ou moins réaliste sont commercialisés depuis 1972 avec l'introduction de la tête KEMAR (figure I.9a). Quelques exemples sont proposés en figure I.9.

Une troisième méthode peut être citée, consistant à effectuer la "binauralisation" lors du rendu par un traitement adapté de signaux.

I.4.2.b Synthèse binaurale

Il est possible de "binauraliser" un signal sonore en lui appliquant l'ensemble des indices de localisation sonore contenus dans les HRTF, comme il a été décrit en I.2.2.

En théorie, une base de données des HRTF complète est composée de l'ensemble des paires de filtres pour toutes les directions de l'espace (r, θ, ϕ). En réalité, une discrétisation de l'espace est nécessaire et le nombre de directions est fini.

Chaque paire de HRTF est associée à une direction donnée. En effectuant le produit de convolution d'un signal monophonique avec les filtres contenus dans cette paire de HRTF, on obtient alors une paire de signaux sonores donnant l'illusion d'une source sonore qui est perçue à la direction déterminée par la paire de filtres, de la même manière que lors de l'écoute naturelle.

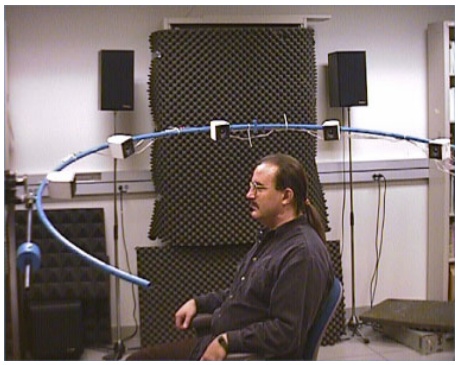
Ainsi, dans le domaine fréquentiel, un signal monophonique S_0 est multiplié par les HRTF gauche (l) et droite (r) pour une direction $(\hat{r}, \hat{\theta}, \hat{\phi})$ donnant une relation équivalente à l'équation (I.3).

Dans le cas où la direction où l'on souhaite placer la source n'est pas disponible dans la base de HRTF, il est possible d'effectuer une interpolation des directions existantes afin d'obtenir la direction souhaitée.

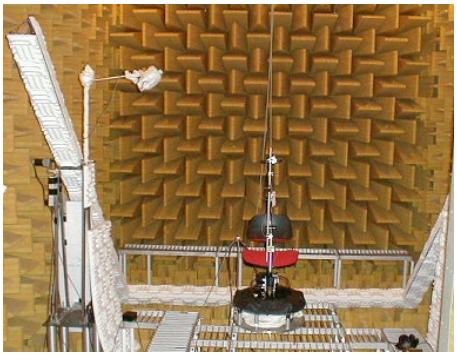
I.4.2.c Les HRTF

Les HRTF étant dépendantes de la morphologie, elles varient fortement d'un individu à un autre. Il est possible de les obtenir par la modélisation ou la mesure.

Les modélisations des HRTF sont effectuées à partir d'un maillage morphologique avec l'utilisation de méthodes numériques telles que les éléments finis [Seppälä et al., 2006] et les éléments de frontière [Katz, 2001]. Ces méthodes demandant un maillage fin, notamment au niveau des oreilles. Malheureusement, les méthodes d'acquisition optique se heurtent à des limitations, notamment lors du scan des parties se trouvant dans des zones d'ombre [Reichinger et al., 2013]. Des techniques plus lourdes comme l'IRM permettent de s'affranchir de ces limitations, mais apportent des contraintes supplémentaires [Otani et al., 2010]. Une fois les maillages obtenus, une phase de post-traitement est obligatoire afin d'éliminer les erreurs de mesure pouvant avoir une incidence sur la modélisation [Ziegelwanger et al., 2013]. Aujourd'hui, ce post-traitement est effectué de façon manuelle constituant le frein principal à cette technique.

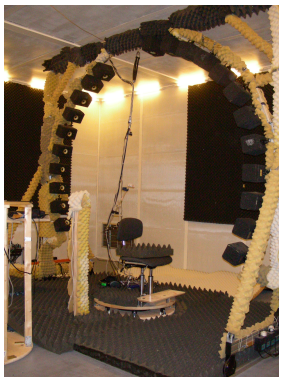


(a) CIPIC [Algazi et al., 2001].



(b) Listen [IRCAM, 2014].

(c) TNO [Pernaux, 2003].



(d) ARI [Austrian Academy of Sciences, 2014].

Figure I.10 : Exemples de dispositifs de mesure de HRTF.

La mesure des HRTF est généralement effectuée dans un environnement anéchoïque. Le sujet est équipé de microphones au niveau du conduit auditif. Ces capteurs peuvent être disposés de façon à obstruer le conduit auditif (méthode "conduit bloqué") ou à l'aide d'une sonde acoustique permettant l'acquisition du signal à l'intérieur du conduit (méthode "conduit ouvert"). La première méthode est généralement privilégiée, car elle permet de s'affranchir des modes engendrés par le conduit, offrant ainsi une meilleure reproductibilité des mesures. Aujourd'hui, avec la miniaturisation de composants électroniques, des nouvelles pistes sont explorées afin d'effectuer des mesures d'intensité au plus près du tympan, mais les rapports signal à bruit ne sont pas encore satisfaisants [Hiipakka, 2013]. Des mesures ont été également effectuées avec des microphones placés au-dessus des oreilles [Austrian Academy of Sciences, 2014] pour les besoins croissants de l'utilisation

du binaural pour les prothèses auditives.

Les sources sonores sont placées de façon équidistante du centre de la tête du sujet. Elles peuvent être disposées de façon fixe ou être déplacées au cours de la mesure à l'aide d'un dispositif mécanique motorisé [Pernaux, 2003].

Les signaux utilisés pour les mesures sont de natures diverses, tels que des *Time Streched Pulse* (TPS), *Maximum Length Sequence* (MLS), sinus glissants, etc.

Afin de s'affranchir de la réponse des transducteurs, les HRTF sont normalisées par rapport à la réponse en fréquence de la chaîne électroacoustique. Celle-ci est mesurée à la position du centre de la tête du sujet (sujet absent).

En fonction de la densité du maillage, les mesures sont généralement très longues, obligeant les sujets à rester tout au long de la mesure dans la même position. Afin de réduire le temps de mesure, des nouvelles méthodes permettent d'effectuer des mesures sur des différentes positions de façon simultanée [Majdak et al., 2007] ou par balayage [Enzner et al., 2011].

Au final, les deux méthodes d'estimation des HRTF présentent des contraintes liées à la mesure et au traitement des données, rendant difficile l'obtention de bases de HRTF personnalisées.

Aujourd'hui, il existe plusieurs bases de données de HRTF dont quelques-unes sont disponibles en libre accès. Chacune présente un maillage ainsi que des conditions de mesurage propres [IRCAM, 2014, Algazi et al., 2001, Grassi et al., 2003, University, 2001, University, 2014, Wierstorf et al., 2011], et certaines sont associées à des maillages morphologiques des sujets [Austrian Academy of Sciences, 2014, Pernaux, 2003, Harder et al., 2013, Guillon et al., 2012].

I.4.2.d Restitution de contenus binauraux

La restitution sur casque stéréophonique (figure I.11) fournit les conditions privilégiées d'écoute de cette technologie, car elle permet d'appliquer directement le signal désiré au plus près du canal auditif.

Cette restitution pose cependant de nombreux problèmes. En premier lieu, le choix du casque et sa calibration sont déterminants. Afin de bien restituer correctement le signal sonore, il est en effet nécessaire de rendre la chaîne de diffusion acoustiquement transparente, en annulant la réponse en fréquence apportée par le casque à l'aide de la mesure des *Headphone Transfert Functions* (HPTF). Ces dernières sont difficiles à évaluer, car elles dépendent de l'individu et du positionnement du casque, qui peut varier entre deux mesures successives [Kulkarni and Colburn, 2000]. Ensuite, les HPTF pré-

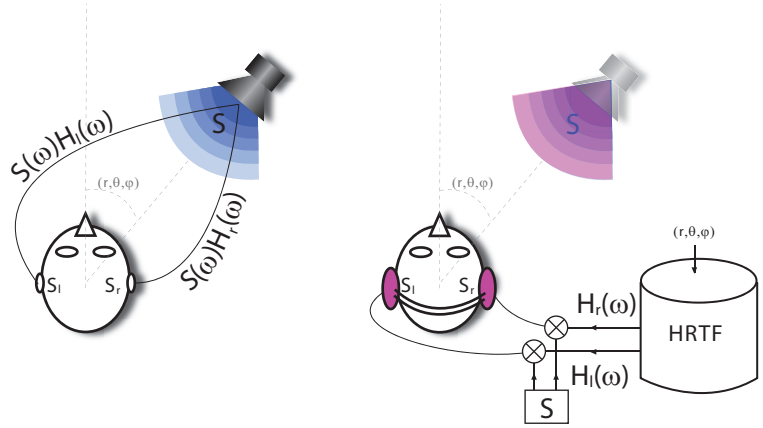


Figure I.11 : Principe de la synthèse binaurale.

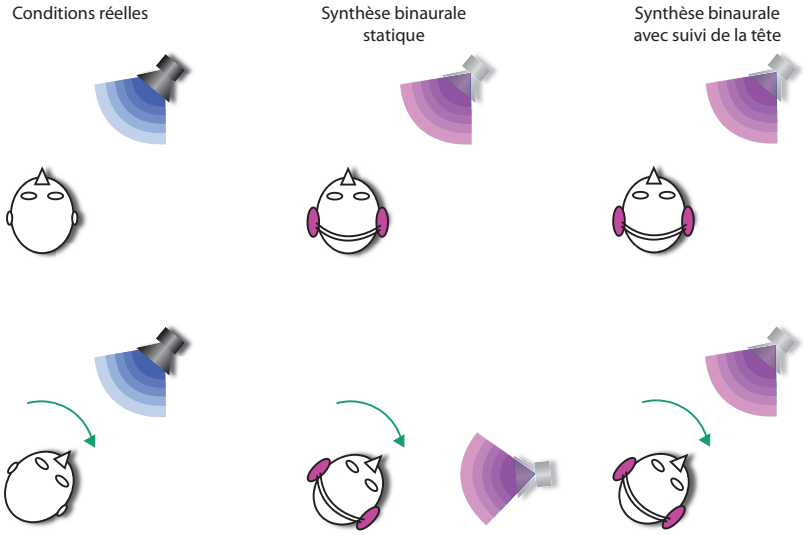


Figure I.12 : Effet de la rotation de la tête en synthèse binaurale statique et dynamique (avec suivi des mouvements de tête).

sentent des altérations similaires aux caractéristiques spectrales des HRTF qui dégradent potentiellement les indices spectraux de localisation. D'autre part, l'occlusion de l'oreille par un casque modifie l'impédance acoustique au niveau du conduit auditif, ce qui altère les conditions d'écoute naturelle. Il est donc proposé d'utiliser des casques de type ouvert pour améliorer la qualité du rendu binaural [Möller, 1992].

Enfin, la synthèse binaurale statique présente l'inconvénient de ne pas pouvoir s'adapter aux mouvements de la tête de l'auditeur. Ce problème est également rencontré lors d'une captation binaurale. La scène sonore suit les mouvements de la tête, s'écartant de conditions réelles et détruisant l'immersion (figure I.12). Dans les conditions réelles, la scène sonore reste statique quelle que soit la position de la tête de l'auditeur, permettant ainsi une localisation fine des sources sonores grâce à l'utilisation d'indices dynamiques

de localisation (I.2.3). Cet inconvénient a été contourné avec la synthèse binaurale dynamique. Dans ce mode de restitution, les mouvements de la tête sont annulés, ce qui implique un suivi de la tête (*head tracking*) en temps réel pour corriger en continu le rendu binaural. Le suivi peut être effectué grâce à des capteurs de déplacement, rotation ou position, couplés à la tête de l'auditeur, ou par un suivi optique. L'implémentation de ces solutions ajoute une latence au rendu sonore. Cette dernière doit être minimale pour ne pas dégrader l'expérience de l'utilisateur. Le *head tracking* augmente la capacité à localiser les sources, notamment en réduisant les confusions avant-arrière [Wenzel, 1995, Wenzel, 1999, Begault et al., 2001, Faure, 2005].

Même si la restitution du binaurale s'effectue de préférence au casque, il est également possible de **restituer des signaux binauraux sur des haut-parleurs**. Si les signaux binauraux sont présentés directement sur des hauts-parleurs, le signal devant être perçu par l'oreille droite est également perçu par l'oreille gauche et inversement, détruisant ainsi l'illusion apportée par le binaural. Il est alors possible d'annuler les trajets croisés avec la technique décrite par Gardner dite *crosstalk cancellation* [Gardner, 1997]. Le signal parasite est ajouté en opposition de phase sur la deuxième voie dans le but de l'annuler. Cette technique permet, en plus d'annuler les trajets croisés, de corriger la réponse des haut-parleurs.

I.4.2.e Pré-traitement des HRTF pour la synthèse binaurale

Lors de l'implémentation des HRTF pour la synthèse binaurale, quelques pré-traitements sont souvent utilisés.

Filtre à phase minimale + retard pur : Une HRTF définit un filtre causal et stable. Il peut donc être divisé en une composante à phase minimale $H_{phMin}(f)$ et un excès de phase $H_{ExPh}(f)$.

Les variations fines de phase perdues lors de la modélisation d'un retard pur ne semblent pas contribuer à la perception de la localisation de sources [Minnaar et al., 1999, Kulkarni et al., 1999].

Le filtre à phase minimale $H_{phMin}(f)$ est calculé de la façon suivante [Kistler and Wightman, 1992, Kulkarni et al., 1995] :

$$H_{phMin}(f) = \begin{cases} |H_{phMin}(f)| = |H(f)| \\ \angle H_{phMin}(f) = \Im[\mathcal{H}(-\log(H(f)))] \end{cases}, \tag{I.4}$$

où $\mathcal{H}$ désigne la transformée de Hilbert.

Le retard à appliquer en complément de la composante à phase minimale peut être calculé soit par un modèle similaire à celui décrit en I.2.1.a, soit par la mesure, dérivé des HRTF utilisant des méthodes comme

- la corrélation en sous-bandes selon la méthode de Wightman et Kristler [Wightman and Kistler, 1992],
- l'enveloppe temporelle proposé par Daniel [Daniel, 2001]
- l'approximation linéaire effectuant une régression linéaire Lin sur l'excès de phase gauche et droite [Jot et al., 1995], suivant

$$ITD = Lin(H_{ExPh,L}(f) - H_{ExPh,R}(f)) \,. \tag{I.5}$$

D'après Larcher [Larcher, 2001], la dernière de ces trois méthodes fournit les meilleurs résultats.

Parmi les méthodes d'estimation du retard étudiées par Nicol [Nicol, 2010], les quatre principales sont :

- le retard de phase moyen dans les basses fréquences de Kulkarni [Kulkarni et al., 1999],
- l'estimation de la pente de la composante à excès de phase de Minnaar [Minnaar et al., 2000],
- l'identification du maximum de corrélation de Nam [Nam et al., 2008],
- l'estimation du retard de groupe moyen proposée par le même auteur [Nam et al., 2008].

Cette étude a démontré que les différentes méthodes sont fortement dépendantes de la base de données de HRTF.

Par la suite, nous avons donc privilégié la première méthode de Nam [Nam et al., 2008], car cette méthode peut être considérée [Nicol, 1999] comme la mieux adaptée à la base de données de Pernaux [Pernaux, 2003], base privée d'Orange et utilisée dans la suite de nos travaux.

Cette méthode consiste à chercher le maximum de la fonction d'inter-corrélation entre les HRIR des deux oreilles.

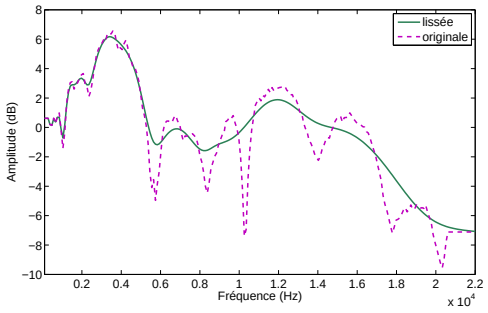


Figure I.13 : HRTF avant et après un lissage utilisant un banc de filtres par bandes critiques ERB.

Lissage spectral : Afin de prendre en compte les limitation de la perception, il est possible d'effectuer un lissage sur l'amplitude du spectre des HRTF (figure I.13). En effet, l'organisation tonotopique [4] de l'oreille interne permet de modéliser le spectre perçu grâce à un banc de filtres représentant chaque région de la membrane basilaire. Il a été démontré dans plusieurs études que la suppression des détails des HRTF n'altère pas les performances de localisation, permettant une diminution de la résolution fréquentielle plus sévère que les filtres auditifs [Guillon, 2009, Huopaniemi and Smith, 1999, Huopaniemi et al., 1999].

Plusieurs méthodes de lissage fréquentiel ont été proposées, soit par un filtrage par banc de filtres [Asano et al., 1990, Kulkarni and Colburn, 1998, Langendijk and Bronkhorst, 2002, Breebaart and Kohlrausch, 2001], soit au moyen de la troncature d'une projection sur une base de vecteurs propres [Kistler and Wightman, 1992, Faure, 2005, Faure et al., 2007, Hacihabiboglu et al., 2002].

Interpolation des HRTF Actuellement, les bases de données des HRTF disponibles (I.4.2.c) ont été mesurées suivant un échantillonnage spatial limité et dans la plupart des cas, uniquement sur la partie supérieure de la sphère ou sur une partie de l'hémisphère inférieur.

Généralement, l'échantillonnage est distribué uniformément sur une coordonnée unique (azimut ou élévation), mais les points de mesure sont très rarement distribués de façon homogène sur l'ensemble de la sphère.

Les directions nécessaires pour la synthèse binaurale ne sont donc pas toujours disponibles dans la discrétisation spatiale utilisée pour l'acquisition des HRTF, notamment dans un but de synthèse dynamique. Afin de rendre les bases de HRTF compatibles avec la synthèse, une interpolation spatiale est alors parfois nécessaire.

Lorsque l'interpolation locale est nécessaire, Larcher [Larcher, 2001] a démontré qu'une

[4] Chaque région de la membrane basilaire est sensible à une certaine plage de fréquences et peut être assimilée à un filtre passe-bande [Guillon, 2009].

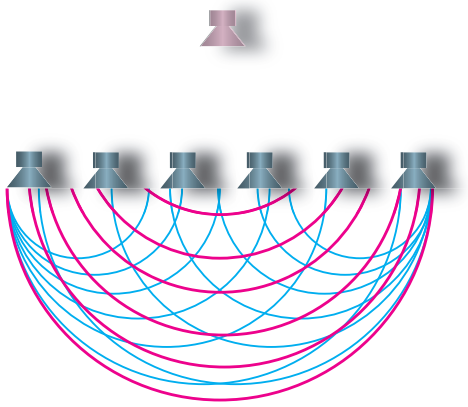


Figure I.14 : Principe de la WFS.

interpolation linéaire des coefficients des filtres à réponse infinie (FIR) donne de bons résultats. Guillon [Guillon, 2009] a montré que, lorsqu'il est nécessaire d'interpoler sur une zone étendue, l'utilisation de la méthode des "Splines sphériques pour coques minces" (*Spherical Thin Plate Spline* (STPS)), tirée de la méthode des "Spline sphériques" de Wahba [Wahba, 1981], est à privilégier, conformément aux résultats de [Hartung and Jonas, 1999].

I.4.3 Méthodes physiques

Les formalismes mathématiques ont permis de proposer des procédés capables de reconstruire le champ acoustique, afin d'immerger l'auditeur dans un champ sonore synthétique proche du champ sonore réel. Nous décrivons ici deux méthodes principales, l'holophonie ou *Wave Field Synthesis* (WFS) et l'ambisonique.

I.4.3.a Holophonie ou WFS

Même si les prémices de la stéréophonie laissaient envisager ce qui serait plus tard l'holophonie (I.3.1.b), ce n'est qu'en 1973 que le terme d'holophonie a été proposé par Jessel pour la première fois [Jessel, 1973]. Cette méthode essaie de recréer le champ acoustique dans un espace donné grâce au principe de Huygens, connu depuis le XVII^{e} siècle. Huygens a démontré que le front d'onde rayonné par une source se comporte comme une distribution d'une infinité de sources secondaires (placées sur ce front d'onde). Selon cette méthode, il est possible de représenter une source quelconque à l'aide de sources secondaires (figure I.14).

Ce principe est généralisé pour exprimer un champ acoustique sur le domaine Ω_1, engendré par des sources monochromatiques de pulsation ω, placées dans le domaine Ω_2 à l'aide de sources placées sur la frontière Ω_0 délimitant les deux domaines. Pour ce faire, il faut exprimer le champ de pression à l'aide de l'intégrale de Kirchhoff [Bruneau, 1998],

$$p(\vec{r},t) = \iint_{\delta\Omega} \left[g(\vec{r}-\vec{r_0},\omega)\vec{\nabla}_0 p_0(\vec{r_0},\omega) - p_0(\vec{r_0},\omega)\vec{\nabla}_0 g(\vec{r}-\vec{r_0},\omega) \right] \vec{n}\mathrm{d}S_0 \,, \tag{I.6}$$

où $\vec{r_0}$ désigne la variable d'intégration, $\vec{n}$ représente le vecteur unitaire normal à la surface $\delta\Omega$, l'indice 0 dénotant le signal sur la frontière Ω_0 et la fonction g représentant la fonction de Green associée au problème. Autrement dit, dans le domaine Ω_1, le champ acoustique est engendré par une infinité de sources élémentaires placées à la frontière Ω_0, dont les propriétés de propagation sont décrites par les termes $g(\vec{r}-\vec{r_0},t-t_0)$ et $\vec{\nabla}_0 g(\vec{r}-\vec{r_0},t-t_0)$, et dont l'amplitude dépend des sources sonores que l'on cherche à représenter.

Généralement, la fonction de Green choisie est la fonction de Green en champ libre,

$$g(\vec{r}-\vec{r_0},\omega) = \frac{e^{ik|\vec{r}-\vec{r_0}|}}{4\pi|\vec{r}-\vec{r_0}|} \,. \tag{I.7}$$

Cette fonction conduit à la décomposition du champ sonore sur l'interface Ω_0 comme une infinité de sources monopolaires associées au même nombre de sources dipolaires, ce qui revient à décomposer le signal acoustique en pression et vitesse acoustique respectivement.

En pratique, seules les sources monopolaires peuvent être conservées. Un haut-parleur monté sur une enceinte close peut être considéré comme une source monopolaire sur une gamme limitée de fréquences. Il est donc possible de créer un champ acoustique à l'aide d'un réseau de haut-parleurs [Nicol, 1999]. Cette approximation limite l'action des interférences destructives créées par les sources dipolaires. Il est donc nécessaire de désactiver les haut-parleurs se trouvant dans la direction opposée à la direction de la source qu'on cherche à reconstituer, afin de recréer artificiellement ces interférences [Nicol, 2010]. Il est également possible de configurer le réseau de haut-parleurs en fonction du champ sonore que l'on cherche à recréer. Par exemple, pour la reconstitution acoustique d'une pièce de théâtre pour un auditeur se trouvant dans un fauteuil dans la salle, tous les haut-parleurs doivent être placés devant la scène, car *a priori* aucune source ne se trouve à côté ou derrière l'auditeur. Comme la distribution des haut-parleurs est dépendante de la scène sonore à représenter, il est difficile de définir une configuration standard, limitant ainsi l'universalité de cette approche.

Le champ acoustique généré peut être virtuel ou réel. Aujourd'hui, il n'existe pas de véritables systèmes de prise de son WFS. En théorie, lorsqu'il s'agit d'un champ acoustique réel, un système de prise de son peut être associé au dispositif de restitution, mais un problème d'échantillonnage spatial se pose. Physiquement, la discrétisation s'impose à cause de la taille des transducteurs et du grand nombre de signaux que cela implique. Cette discrétisation de l'espace génère alors un repliement spatial qui se produit à partir de la fréquence $f_{al} = \frac{c}{2\Delta}$ pour un réseau circulaire horizontal de transducteurs homogènement distribués et écartés de la distance Δ. De plus, le principe énoncé par l'équation (I.6) n'est valable que si le champ acoustique est recréé par une infinité de sources secondaires.

I.4.3.b Ambisonique

De la même manière que la WFS, cette technique cherche à caractériser le champ acoustique à travers une décomposition spatiale Le format B de l'ambisonique, tel que défini pour la première fois en 1973 par Gerzon [Gerzon, 1973a], est la décomposition d'une onde acoustique $p(kr, \theta, \phi)$ [5] de nombre d'onde $k = \frac{\omega}{c}$ sur les ordres 0 et 1 des harmoniques sphériques Y_{mn}^{σ} [Potel and Bruneau, 2006, Groemer, 1996].

L'ordre 0 ou composante "W" peut être interprété comme le signal de pression acoustique, et les trois composantes du premier ordre "X", "Y" et "Z" correspondent aux composantes du vecteur définissant la vitesse acoustique particulaire. La méthode définie par Gerzon n'est qu'un cas particulier de la projection des ondes acoustiques sur la base d'harmoniques sphériques.

Harmoniques sphériques : En analogie avec les séries de Fourier, toute fonction f définie sur la sphère peut s'écrire comme un développement en séries d'harmoniques sphériques (annexe C),

$$f(\theta, \phi) = \sum_{m=0}^{\infty} \sum_{n=0}^{m} \sum_{\sigma=\pm 1} f_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \phi) \,, \tag{I.8}$$

où

[5] Exprimée en coordonnées sphériques r, θ, ϕ où r définit la distance, θ l'azimut et ϕ l'élévation.

$$\begin{aligned} f_{mn}^{\sigma} &= \langle f, Y_{mn}^{\sigma} \rangle_S \\ &= \frac{1}{4\pi} \int_{\theta=0}^{2\pi} \int_{\phi=\frac{-\pi}{2}}^{\frac{\pi}{2}} f(\theta,\phi) Y_{mn}^{\sigma}(\theta,\phi) \cos\phi \, \mathrm{d}\phi \, \mathrm{d}\theta \, , \end{aligned} \tag{I.9}$$

$\langle f, g \rangle_S$ exprimant le produit scalaire de deux fonctions f et g définies sur la sphère S. Les harmoniques sphériques Y_{mn}^{σ} sont des fonctions spatiales de θ et de ϕ définis par

$$Y_{mn}^{\sigma}(\theta,\phi) = \sqrt{(2m+1)\epsilon_n \frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\theta) \times \begin{cases} \cos n\theta, & \text{si } \sigma = 1 \\ \sin n\theta, & \text{si } \sigma = -1 \end{cases} , \tag{I.10}$$

où

$$\begin{cases} m, n \in \mathbb{N}, n \leq m, \\ \sigma \in \{-1, 1\}, \\ \epsilon_0 = 1, \text{ et } \epsilon_n = 2\text{si}n > 0. \end{cases}$$

m étant l'ordre de l'harmonique et P_{mn} la fonction de Legendre associée, celle-ci étant définie pour $x \in [-1, 1]$ par

$$P_{mn}(x) = (1 - x^2)^{\frac{n}{2}} \frac{d^n}{dx^n} P_m(x) \, , \tag{I.11}$$

où $P_m(x)$ est le polynôme de Legendre de première espèce tel que

$$\begin{cases} P_0(x) = 1 \\ P_1(x) = x \\ (m+1)P_{m+1}(x) = (2+1)xP_m(x) - mP_{m-1}(x), m > 1 \end{cases} . \tag{I.12}$$

Dans le cas de l'acoustique linéaire, une onde acoustique monochromatique de nombre d'onde $k = \frac{\omega}{c}$ de pulsation ω et se propageant à une vitesse c obéit à l'équation de

Helmholtz [Bruneau, 1998, Potel and Bruneau, 2006]

$$(\Delta + k^2)p(r, \theta, \phi) = 0\,, \tag{I.13}$$

où p désigne la pression acoustique. En absence de source, cette équation trouve une solution générale à l'intérieur de la sphère de rayon R ($r < R$) appelée *série de Fourier-Bessel* [Bruneau, 1998]

$$p(kr, \theta, \phi) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \phi), \tag{I.14}$$

où les fonctions de dépendance radiale $j_m(kr)$ sont les *fonctions de Bessel sphériques* définies par

$$\begin{cases} j_0 = 1 \\ j_m = (-1)^m x^m \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^m \frac{\sin x}{x} \end{cases}, \tag{I.15}$$

où les coefficients B_{mn}^{σ} sont obtenus par projection orthogonale de l'expression de la pression acoustique p sur les harmoniques sphériques Y_{mn}^{σ} conformément à l'équation I.9,

$$\begin{aligned} i^m j_m(kr) B_{mn}^{\sigma} &= \langle p, Y_{mn}^{\sigma} \rangle_S \\ &= \frac{1}{4\pi} \int_{\theta=0}^{2\pi} \int_{\phi=\frac{-\pi}{2}}^{\frac{\pi}{2}} p(kr, \theta, \phi) Y_{mn}^{\sigma}(\theta, \phi) \cos\phi \, \mathrm{d}\phi \, \mathrm{d}\theta\,. \end{aligned} \tag{I.16}$$

La projection sur une infinité d'harmoniques sphériques n'est possible qu'en théorie et une troncature à un ordre $M \in \mathbb{N}$ s'impose. L'équation (I.14) devient alors

$$p(kr, \theta, \phi) = \sum_{m=0}^{M} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \phi)\,. \tag{I.17}$$

Cette approximation donne un nombre de K coefficients B_{mn}^{σ} tel que

$$K = (M+1)^2 \,. \tag{I.18}$$

Lorsque l'ordre M de troncature est égal à 1, on obtient la représentation proposée par Gerzon comme le format B. Lorsque M est supérieur à 1, on parle alors de l'ambisonique aux ordres supérieurs ou *Higher Order Ambisonics* (HOA) [Bamford, 1995] [Daniel, 2001].

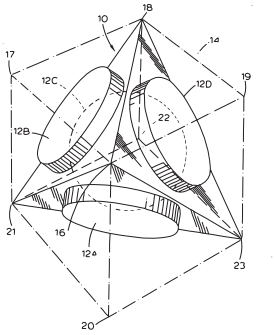
I.4.3.c Ambisonique à l'ordre 1 et aux ordres supérieurs

Tel que décrit précédemment, le format B de l'ambisonique est la projection d'une onde acoustique $p(kr, \theta, \phi)$ sur les deux premiers ordres des harmoniques sphériques Y_{mn}^{σ}. La composante d'ordre zéro (également connue comme la composante "W") représente la pression acoustique, et les composantes du premier ordre (également connues sous le nom de "X", "Y" et "Z" respectivement) correspondent aux composantes du vecteur vitesse selon les directions x, y et z d'un espace 3D en coordonnées cartésiennes. Cette représentation est une troncature à l'ordre 1 de la relation (I.17) et est exprimée sous la forme :

$$\begin{array}{rccc} p(kr,\theta,\phi) = & \frac{\sin kr}{kr} & B_{00}^{1}(\theta,\phi) & 1 \\ +i & \frac{\sin kr}{kr^2} - \frac{\cos kr}{kr} & B_{11}^{1}(\theta,\phi) & \cos(\phi)\cos(\theta) \\ +i & \frac{\sin kr}{kr^2} - \frac{\cos kr}{kr} & B_{11}^{-1}(\theta,\phi) & \cos(\phi)\sin(\theta) \\ +i & \underbrace{\frac{\sin kr}{kr^2} - \frac{\cos kr}{kr}}_{j_m(kr)} & \underbrace{B_{10}^{1}(\theta,\phi)}_{B_{mn}^{\sigma}} & \underbrace{\sin(\phi)}_{Y_{mn}^{\sigma}} \end{array} \,. \tag{I.19}$$

Les signaux "W", "X", "Y" et "Z" du format B de l'ambisonique sont obtenus à la prise de son à l'aide d'un microphone ambisonique communément appelé microphone Soundfield® (figure I.15b).

Ce dispositif défini par Gerzon en [Gerzon, 1975] est composé de quatre capsules microphoniques cardioïdes montées sur les faces d'un tétraèdre (figure I.15a). Un matriçage permet le passage des signaux issus des capteurs ("format A") vers le "format B" correspondant aux signaux W,X,Y,Z et définissant la représentation ambisonique de l'onde acoustique. Daniel [Daniel, 2001] a démontré d'une part, qu'en plaçant les capsules microphoniques aux sommets des polyèdres réguliers, il est possible d'atteindre le deuxième ordre, et, d'autre part, que pour étendre à des ordres plus élevés, il est nécessaire d'utiliser des polyèdres semiréguliers tels que celui à 32 sommets. Cette configuration est maintenant utilisée par le microphone Eigenmike de la société *mh acoustics* illustre en figure I.15c.



(a) Schèma du brevet de Gerzon [Gerzon and Craven, 1977].

(b) Soundfield SPS200.

(c) Eigenmike de mh acoustics.

Figure I.15 : Microphones ambisoniques.

Lors du passage des signaux microphoniques ("format A") vers les signaux ambisoniques ("format B"), une correction de distance physique des capsules par rapport au centre du dispositif doit être effectuée [Gerzon and Craven, 1977]. Les signaux obtenus doivent êtres normalisés par rapport à la pression acoustique. La pondération appliquée dépend de la convention choisie. En effet, les différents auteurs travaillant dans le domaine ont proposé des conventions différentes (tableau I.2). Une attention particulière doit donc être portée à cette pondération afin de ne pas introduire un biais dans les signaux ambisoniques.

Signal	Gerzon*	Furse et Malham**	Bamford***
W	1,	$1/\sqrt{2}$	1
X	$\sqrt{2}\cos\theta\cos\phi$,	$\cos\theta\cos\phi$,	$\cos\theta\cos\phi$,
Y	$\sqrt{2}\sin\theta\cos\phi$,	$\sin\theta\cos\phi$,	$\sin\theta\cos\phi$,
Z	$\sqrt{2}\sin\phi$,	$\sin\phi$,	$\sin\phi$.

Tableau I.2 : Différentes conventions de normalisation des composantes ambisoniques (*[Gerzon, 1973b], **[Furse, 2014, Furse and Malham, 2005, Malham, 2008], ***[Bamford, 1995])

Cas d'une représentation bi-dimensionnelle : Très souvent, le champ sonore que l'on cherche à capter se trouve majoritairement dans un seul plan (par exemple, dans le cas de la téléconférence ou un programme radio, l'ensemble de locuteurs se trouvent sur le plan horizontal). De même, le système de restitution du champ acoustique est souvent composé d'un ensemble de haut-parleurs placés sur le plan horizontal (Système 5.1 par exemple). Dans ces cas, le champ sonore est décrit uniquement par les coordonnées (r, θ), réduisant ainsi le nombre des signaux ambisoniques le décrivant.

La décomposition du champ acoustique se fait alors sur une base d'harmoniques cylindriques Y_m^σ définis par :

$$\begin{cases} Y_0^1 = 1, \\ Y_m^\sigma = \sqrt{2}\cos(m\theta) \text{ si } \sigma = 1, \\ Y_m^\sigma = \sqrt{2}\sin(m\theta) \text{ si } \sigma = -1, \end{cases} . \tag{I.20}$$

Dans cette description, le nombre de signaux ambisoniques est fortement réduit a

$$K = 2M + 1 \, . \tag{I.21}$$

Marschall [Marschall and Chang, 2013] tire avantage de cette représentation dans une approche d'ordre hybride définissant la pression acoustique sous la forme

$$\begin{aligned} p(kr, \theta, \phi) = &\sum_{m=0}^{M_{3D}} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^\sigma Y_{mn}^\sigma(\theta, \phi) \\ &+ \sum_{m=M_{3D}+1}^{M_{2D}} i^m j_m(kr) \sum_{\sigma=\pm 1} B_{mn}^\sigma Y_{mn}^\sigma(\theta, \phi) \, , \end{aligned} \tag{I.22}$$

où M_{2D} correspond à l'ordre de troncature pour une représentation 2D et M_{3D} à celui d'une répresentation 3D. Cette méthode traite le champ horizontal dans une approche 2D et le reste de la sphère dans une approche 3D, ce qui permet d'adapter la précision de la restitution au flou perceptif selon chaque cordonnée.

La restitution de l'ambisonique est effectuée par un réseau sphérique de L haut-parleurs distribués sur la sphère 3D. Chaque composante ambisonique (B^{σ}_{mn}) est exprimée comme une somme pondérée des signaux émis par les haut-parleurs (S_l). Les haut-parleurs sont définis comme des sources spatialisées dans la base d'harmoniques sphériques Y_{mn}, permettant ainsi d'en déduire les coefficients de pondération qui leur sont associés. La reconstruction (ou transcodage) des composantes HOA sur des haut-parleurs peut être obtenue grâce à la relation

$$b = C \cdot s \,, \tag{I.23}$$

où

$$C = \begin{pmatrix} Y^1_{00}(\theta_1,\phi_1) & \cdots & Y^1_{00}(\theta_L,\phi_L) \\ Y^1_{11}(\theta_1,\phi_1) & \cdots & Y^1_{11}(\theta_L,\phi_L) \\ Y^{-1}_{11}(\theta_1,\phi_1) & \cdots & Y^{-1}_{11}(\theta_L,\phi_L) \\ Y^1_{10}(\theta_1,\phi_1) & \cdots & Y^1_{10}(\theta_L,\phi_L) \end{pmatrix},$$
$$s = \begin{pmatrix} S_1 \\ \vdots \\ S_L \end{pmatrix}, \; b = \begin{pmatrix} B^1_{00} \\ B^1_{11} \\ B^{-1}_{11} \\ B^1_{10} \end{pmatrix}.$$

S_l correspond au signal émis par le haut-parleur l situé dans la direction (θ_l, ϕ_l). La matrice C contient les harmoniques sphériques associés à la position des haut-parleurs. Pour que cette équation puisse être résolue, la reproduction doit être effectuée sur un jeu composé d'au moins quatre haut-parleurs pour le décodage de l'ambisonique du premier ordre, tel que défini par la relation I.18.

L'équation (I.23) est obtenue en faisant correspondre les coefficients B^{σ}_{mn} du champ reconstruit avec ceux de l'onde acoustique décrite par l'équation (I.19). La résolution de l'équation (I.23) permet l'obtention des signaux diffusés par les haut-parleurs à partir des signaux ambisoniques à l'aide de la matrice de décodage D

$$s = D \cdot b \,, \tag{I.24}$$

où la matrice D est la matrice inverse de la matrice C. Elle peut être obtenue grâce à la

méthode de la matrice pseudo-inverse de Moore-Penrose [Golub and Loan, 1996]

$$D = C^t \cdot (C \cdot C^t)^{-1} \,. \tag{I.25}$$

Si les L haut-parleurs formant l'ensemble de restitution sont uniformément distribués sur la sphère, la relation (I.25) devient [Moreau, 2006]

$$D = \frac{1}{L} C^t \,, \tag{I.26}$$

car

$$C \cdot C^t = \frac{1}{L} I_L \,, \tag{I.27}$$

où I_L est la matrice identité de taille $L \times L$.

I.5 Éventail des outils audio 3D disponibles aujourd'hui

Tout au long des paragraphes précédents, nous avons établi un recueil des technologies disponibles aujourd'hui. Certaines d'entre elles font l'objet de produits déjà présents sur le marché. La figure I.16 en propose une synthèse. Chaque technologie est disposée dans le niveau qu'elle occupe dans la chaîne électroacoustique 3D.

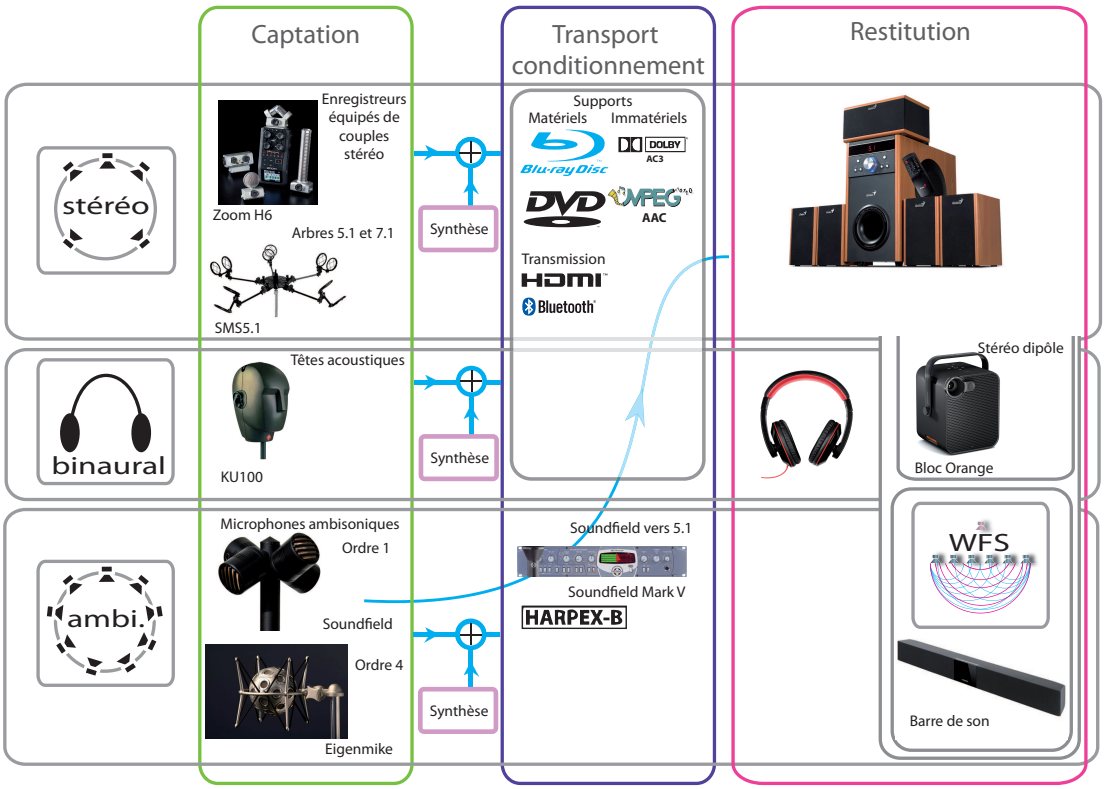


Figure I.16 : Dispositifs disponibles aujourd'hui sur le marché intégrant la chaîne électroacoustique 3D.

I.5.1 Encodage spatial

Nous voulons présenter ici les différents formats existants pour représenter une scène sonore. Ces méthodes peuvent être abordées sous deux aspects. Il est possible de les regrouper d'une part, par rapport aux méthodes utilisées pour la gestion des flux sonores composant la scène sonore, et, d'autre part, par la façon dont la scène sonore est représentée. Les différentes méthodes ont leur place au sein de familles qui sous-divisent ces deux classifications.

La première classification englobe la façon dont les signaux sonores analogiques ou numériques sont encapsulés. Aujourd'hui, il existe une large variété de formats disponibles qui peuvent être regroupés en deux familles :

- la première est basée sur un enregistrement sans perte tel que le *Pulse-Code Modulation* (PCM) où le signal est stocké en sortie du Convertisseur Analogique - Nunérique (CAN),
- la deuxième famille englobe les méthodes d'encodage avec perte qui permettent une compression du signal afin de réduire le volume de données à stocker et à transmettre. Elles comportent généralement une perte de l'information, perte qui peut ou non être audible.

Dans la deuxième classification, nous pouvons intégrer les différentes techniques présentées en I.3.1.c et I.3.2. En effet, elles effectuent une discrétisation spatiale de la scène sonore sur un nombre limité de signaux et paramètres pouvant la représenter. A ce titre, ces méthodes peuvent être considérées comme des formats de représentation spatiale permettant d'effectuer un encodage de la scène sonore (sans compression du point de vue du signal). Nous pouvons dénombrer trois familles principales que nous classifierons sous les noms de méthodes :

- "basée canaux" ou *channel based*,
- "basée sur le champ sonore" ou *soundfield based*,
- "basée objet" ou *Object-based*.

Comme ces deux classifications se recouvrent et sont interdépendantes, nous allons aborder ici l'ensemble des techniques, utilisant le point de vue de la seconde classification. Ces techniques sont donc regroupées suivant la manière dont elles représentent la scène sonore.

I.5.1.a Formats *channel-based*

Les systèmes stéréophoniques décrivent le champ acoustique avec une approche reposant sur les canaux qui le composent. Ces derniers sont directement liés et dépendants du dispositif de restitution, où le nombre de canaux correspond exactement au nombre de haut-parleurs utilisés, la scène sonore étant décrite par un ensemble de signaux ou canaux sonores donnant le nom à cette famille.

Les méthodes *channel based* étant le standard jusqu'à maintenant, de nombreuses stratégies permettant leur acheminement ou compression ont été proposées. Tout d'abord, nous pouvons citer **les techniques de matriçage**. En effet, elles permettent une réduction de nombre de signaux, facilitant ainsi leur transmission et/ou leur stockage. La réduction est effectuée par des opérations mathématiques simples entre les signaux afin d'éliminer l'information redondante. Ce type d'encodage peut être effectué sans perte, avec des méthodes comme la *mid/side stereo coding* [Johnston and Ferreira, 1992] et la *Meridian Lossless Packing* (MLP) [Gerzon et al., 1999] (utilisée dans les DVD et les Blue-ray dans l'encodage AC-3), permettant ainsi une réduction de canaux de 2 :1.

Des techniques plus évoluées permettent le matriçage avec perte [Dressler, 2000]. La technique *Dolby Surround* permettant une compression 4 :2 et la *perceptual mid/side stereo coding* font partie de cette famille. La méthode *perceptual mid/side stereo coding* permet une compression de 2 :1 et est utilisée à la base des méthodes de compression Dolby stéréo, Dolby AC-3 et l'MPEG AAC.

La compression des signaux à large bande a débuté avec les approches perceptives de la norme MPEG et ses couches (*layers*) respectives I, II et III. Cette méthode normalisée permet une réduction considérable du débit des signaux encodés. A partir de la deuxième couche, cet encodage tire avantage de la redondance de l'information entre les deux canaux stéréophoniques. Elle est exploitée selon deux modalités de matriçage : l'encodage de la stéréo d'intensité (*intensity stereo coding*) et le codage stéréophonique MS (*middle/side stereo coding*). La première modalité constitue un signal moyen pour les 32 sous-bandes, et les canaux stéréophoniques sont construits à partir d'un facteur d'échelle du signal moyen. La seconde modalité considère certaines sous-bandes comme une somme (M) ou une soustraction (S) de ces deux canaux [Pan, 1995].

Une technique similaire à celle utilisée par la couche II du MPEG a été introduite par Dolby via son codec AC-2, permettant l'encodage de la stéréophonie multicanale par paires indépendantes, au même titre que la MPEG II. La méthode AC-3 de Dolby a été spécialement conçue pour l'encodage du 5.1 permettant également le *downmix* vers la stéréo [Painter and Spanias, 2000] [Herre et al., 2004] [Daniel et al., 2010].

Aujourd'hui, des nouvelles méthodes de compression voient le jour. Le nombre croissant de canaux de restitution utilisés impose de trouver des nouvelles solutions pour l'encodage de la scène sonore.

Jusqu'à maintenant, des stratégies permettant la combinaison des canaux et l'élimination de redondances entre eux étaient à la base des techniques utilisées. L'**encodage spatial paramétrique** repose en particulier sur une analyse de la scène sonore spatiale afin d'extraire les informations nécessaires à sa description en termes de distribution spatiale du signal sonore. La stéréo paramétrique (*parametric stereo*) [Breebaart et al., 2005] et le*Binaural cue coding* [Faller, 2004] ont démontré la faisabilité de ce type de techniques avec un encodage de descripteurs perceptifs tels que les *Inter-channel Time Difference* (ICTD) et *Inter-channel Level Difference* (ICLD). Ces techniques, basées sur des signaux stéréophoniques, limitent leur décodage à un système de restitution différent de celui d'origine. La méthode adoptée par la norme *MPEG Surround* utilise également ce type de description et permet également un rendu sur un système de restitution différent à celui d'origine.

I.5.1.b Formats *soundfield-based*

Le méthodes auparavant appelées méthodes physiques, notamment les méthodes ambisonique et HOA effectuent une décomposition de la scène sonore sur un ensemble de vecteurs la décrivant.

I.5.1.c Formats *object-based*

Une dernière méthode peut être définie. Il s'agit du "format objet" ou les *Spatial Audio Object Coding* (SAOC). En effet, un espace sonore 3D peut être décrit complètement grâce à une description des objets sonores qui le composent (sources ponctuelles, distribuées ou diffuses), leur position au cours du temps et leur interaction avec l'environnement (réflexions, diffractions, atténuations) permettant ainsi une réduction considérable du débit (figure I.17).

Ces formats présentent un avantage considérable lors du décodage : leur représentation, les rend indépendants de la méthode ou du système de restitution, à condition d'effectuer un décodage adapté [Geier et al., 2010].

Il est possible de considérer les méthodes de codage spatial de la scène sonore *Spatial Audio Scene Coding* (SASC) comme faisant partie de ces formats à condition que les sources puissent être discriminées les unes des autres.

Les méthodes comme le codage audio directionnel SASC [Goodwin and Jot, 2008] et le *Directional Audio Coding* (DirAC)[Pulkki, 2006] proposent une analyse directionnelle de la scène sonore et la décrivent à partir d'un signal sonore et des méta-données spatiales, permettant de définir la position et la nature de la source (ponctuelle ou diffuse) dans une approche temps-fréquence.

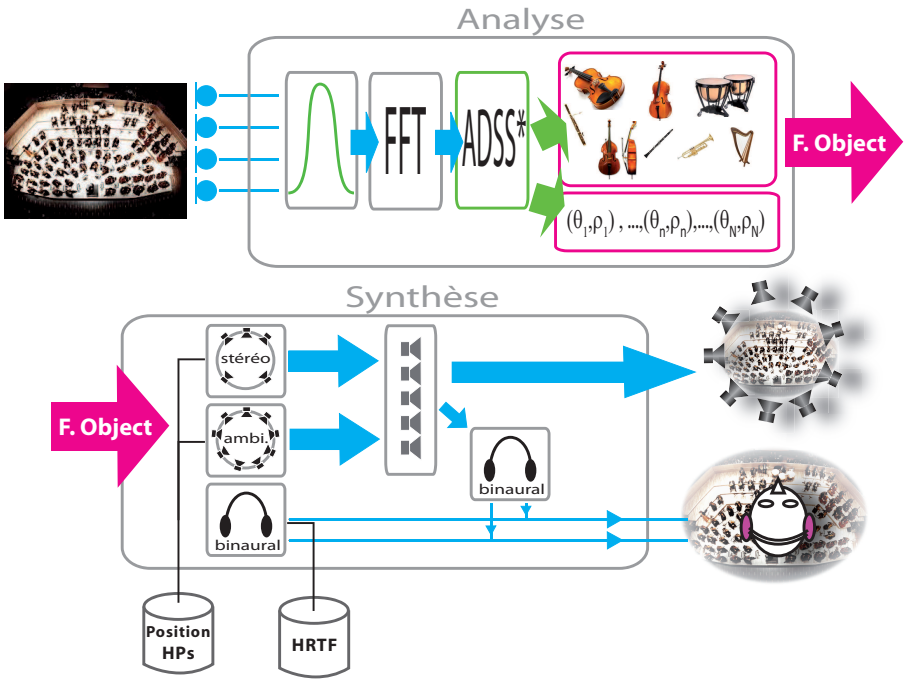


Figure I.17 : Schéma des phases d'analyse et synthèse des méthodes *Object-based* (*Analyse De la Scène Sonore (ADSS)).

La méthode DirAc, détaillée en annexe D, utilise les signaux du format B de l'ambisonique à l'ordre 1 pour extraire les informations liées à la position des sources. En effet, les signaux X,Y,Z étant équivalents aux composantes de la vitesse acoustique, grâce à la relation (D.7), il est possible d'en extraire le vecteur intensité correspondant. Ce vecteur permet, en connaissant sa direction, de déterminer la direction de la source qui le produit, car le vecteur intensité est colinéaire à la propagation acoustique.

Les nouveaux formats Auro 3D, Dolby Atmos, DTS *Multi Dimensional Audio* (MDA) et la nouvelle norme MPEG-H [Breebaart et al., 2008] se servent de ce type de description spatiale dans un mode d'utilisation privilégié. Ils permettent ainsi une restitution sans perte de qualité grâce à l'utilisation de méta-données qui, sur la base de la position des sources (représentées par des signaux monophoniques indépendants), s'affranchissent du flux sonore qui n'est plus déterminé par le nombre de canaux de restitution.

I.5.2 Le *downmix* binaural : l'outil universel de restitution

Le binaural présente principalement deux avantages :

- sa capacité à reproduire une scène sonore spatialisée sur l'ensemble de l'espace 3D avec uniquement deux signaux,
- utiliser un dispositif de restitution générique tel qu'un casque stéréo pour restituer les signaux à proximité du canal auditif.

Le décodage peut s'appliquer directement à l'aide des méthodes *Channel-based* remplaçant les haut-parleurs la composant par des haut-parleurs virtuels. En revanche, les méthodes *soundfield-based* et *Object-based* nécessitent une phase plus complexe de décodage pour l'adaptation au format binaural.

I.5.2.a Principe

Dans son approche la plus basique, il est possible de restituer une source monophonique dans une direction particulière de la sphère 3D, en filtrant ce signal avec les HRTF correspondant à la direction souhaitée. Dans une approche un peu plus complexe, il est possible de restituer correctement une prise de son stéréophonique (2.0, 5.1 7.1 ou 22.2), en filtrant le signal de chaque haut-parleur par les HRTF correspondant à leur positions. Cette technique donne de très bons résultats, à condition que l'implémentation des filtres soit effectuée soigneusement [McKeeg and McGrath, 1997, Breebaart et al., 2006, Breebaart, 2007, Pike and Melchior, 2013]. La restitution étant effectuée sur casque, les problèmes liés au dispositif d'origine sont écartés.

Par extension, il est possible de restituer tout type de spatialisation sonore conçue pour une diffusion sur haut-parleurs. En particulier, pour la restitution de l'ambisonique, les signaux B^{σ}_{mn} peuvent être décodés sur un ensemble sphérique ou circulaire de haut-parleurs. Les signaux les alimentant sont obtenus par projection des signaux ambisoniques sur les harmoniques sphériques associés à la position des haut-parleurs (I.4.3.b). Dans le but d'un décodage sur casque, les haut-parleurs ne sont pas de dispositifs physiques. On parle alors de haut-parleurs virtuels. Les signaux décodés pour les haut-parleurs virtuels sont donc filtrés par les HRTF associées à leur position, de la même façon que pour les systèmes stéréophoniques.

Restitution binaurale de l'ambisonique Tel que précisé en I.5.2, il est possible de l'effectuer grâce à un décodage adapté. Le décodage binaural est basé sur la diffusion des signaux sonores sur un ensemble de haut-parleurs virtuels qui sont créés par filtrage avec

les HRTF des directions correspondantes. Cette méthode permet de recréer un champ acoustique à l'entrée des oreilles de l'auditeur. La procédure est effectuée à l'aide de la matrice de décodage E, qui permet le passage de signaux ambisoniques contenus dans la matrice b de dimension $N \times 1$, vers la matrice F_{bin} de dimension 2×1 contenant les signaux binauraux,

$$F_{bin}\cdot = E \cdot s \,. \tag{I.28}$$

Les signaux binauraux, $F_{bin,\mathbf{l}}$ et $F_{bin,\mathbf{r}}$, sont obtenus en simulant la propagation de chaque haut-parleur aux deux oreilles de l'auditeur en appliquant les filtres des HRTF correspondants. Ce filtrage peut être effectué grâce à la relation matricielle suivante :

$$F_{bin}\cdot = H_{bin} \cdot S \,, \tag{I.29}$$

où S est le spectre du signal audio s alimentant les haut-parleurs virtuels à la pulsation ω, avec

$$F_{bin}(\omega) = \begin{bmatrix} F_{bin,\mathbf{l}}(\omega) \\ F_{bin,\mathbf{r}}(\omega) \end{bmatrix} \,, \tag{I.30a}$$

$$H_{bin}(\omega) = \begin{bmatrix} H_{\mathbf{l}}(\omega, \theta_1, \phi_1) & \dots & H_{\mathbf{l}}(\omega, \theta_L, \phi_L) \\ H_{\mathbf{r}}(\omega, \theta_1, \phi_1) & \dots & H_{\mathbf{r}}(\omega, \theta_L, \phi_L) \end{bmatrix} \,. \tag{I.30b}$$

$H_{bin}(\omega)$ est donc la matrice qui définit l'ensemble de HRTF mesurées dans les L directions des haut-parleurs. Pour réaliser la *binauralisation* des signaux émis par les haut-parleurs, en remplaçant la matrice S de (I.29) par la transformée de Fourier de son expression définie dans l'équation (I.24), on obtient

$$F_{bin} = H_{bin} \cdot D \cdot b \,, \tag{I.31}$$

et la matrice de décodage E introduite en (I.28) peut être obtenue suivant

$$E = H_{bin} \cdot D. \tag{I.32}$$

Cette procédure est équivalente à effectuer une projection d'un ensemble de HRTF sur une base d'harmoniques sphériques [Larcher, 2001] permettant la restitution binaurale des signaux ambisoniques par un simple produit matriciel.

D'autres stratégies d'adaptation des signaux ambisoniques pour leur restitution sur casque sont possibles, utilisant une première phase de localisation des sources tel que décrit dans les paragraphes suivants.

I.5.2.b Décodage binaural actif : une nouvelle variante exploitant le format *object-based*

Il existe aujourd'hui des méthodes novatrices pour le décodage de l'ambisonique pour une restitution binaurale. Ces techniques partent de l'hypothèse que le champ sonore peut ne pas être restitué de façon parfaite. Une approximation du champ sonore est suffisante pour satisfaire la perception du sonore restitué.

Des hypothèses communes à ces décodages permettent d'extraire uniquement les paramètres décrivant l'interaction entre les propriétés du champ sonore et les caractéristiques perceptives engendrées par celui-ci, à savoir,

- la direction de la source est perçue grâce aux paramètres perceptifs de localisation ITD, ILD et indices spectraux,
- la direction de la source et le champ diffus sont mesurés pour des trames temps-fréquence.

Cette dernière hypothèse résulte du fait que ces méthodes supposent que le système auditif n'est capable de décoder qu'une seule source à chaque instant et pour chaque échantillon fréquentiel.

Ceci permet d'encoder le signal sonore de façon à ne conserver qu'un signal représentatif de la scène sonore et des indices déterminant la direction de la source à reproduire, ces méthodes de *Spatial Audio Coding* (SAC) s'approchant ainsi des méthodes d'encodage SAOC.

Ce type de décodage peut être divisé en deux parties, la partie analyse et la partie synthèse.

Pour chacune des parties, le processus se déroule au travers d'une décomposition temps-fréquence dont la largeur des fenêtres temporelles et fréquentielles est liée à la perception sonore ainsi qu'aux contraintes techniques du dispositif utilisé.

L'analyse et la synthèse s'effectuent de façon indépendante sur chaque trame spectro-temporelle (figure I.17). Dans la première phase (analyse), la direction des sources est déterminée et le signal représentatif est extrait. Des méthodes usuelles de localisation de sources pouvant être utilisées dans cette première phase sont décrites en annexe A. Dans la phase de synthèse, le champ sonore constituant le signal représentatif est restitué dans la direction correspondante, prenant en compte des paramètres perceptifs et les contraintes du dispositif de restitution. Ces méthodes peuvent être considérées comme un transcodage d'une méthode *soundfield based* vers une méthode SAC, rendant les contenus compatibles avec tout dispositif de restitution.

Harpex : Berge et Barret [Berge and Barrett, 2010] ont proposé comme alternative au décodage classique du binaural de l'ambisonique (I.4.3.c) la méthode baptisée Harpex.

Ce décodeur paramétrique est, comme la méthode DirAC, basé sur une analyse de la scène sonore. Tout d'abord, les sources sonores sont localisées pour ensuite effectuer un décodage classique (I.4.3.c) sur les directions obtenues lors de la première phase.

L'analyse de la scène sonore est effectuée de la façon suivante.

A chaque trame temporelle et pour chacune des bandes fréquentielles, la position possible des deux sources sonores est déterminée. L'extraction des directions à partir du format B de l'ambisonique est effectuée par la décomposition du signal sonore en deux ondes planes se propageant le long des axes $\sigma_1 = (X1, Y1, Z1)$ et $\sigma_2 = (X2, Y2, Z2)$, grâce à la relation suivante,

$$\underbrace{\begin{bmatrix} \sqrt{2}W \\ X \\ Y \\ Z \end{bmatrix}}_{\mathrm{B}} = \underbrace{\begin{bmatrix} 1 & 1 \\ X_1 & X_2 \\ Y_1 & Y_2 \\ Z_1 & Z_2 \end{bmatrix}}_{\mathrm{V}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}}_{\mathrm{A}}, \tag{I.33}$$

où V est la matrice des composantes ambisoniques des signaux sonores se propageant selon σ_1 et σ_2, et les composantes du vecteur A déterminent leurs amplitudes. En d'autres termes, pour chaque trame temps-fréquence, deux composantes sont identifiées à partir de l'information spatiale fournie par le format B. Elles sont considérées par approximation comme des ondes planes et leur direction est ensuite utilisée pour un décodage ambisonique.

Dans [Berge and Barrett, 2014], Berge explique que si les signaux ainsi obtenus sont directement filtrés avec les HRTF associées aux directions σ_1 et σ_2, des artefacts audibles altèrent le rendu binaural, particulièrement lorsque les sources se déplacent au cours du temps.

Pour cette raison, le décodage binaural est effectué de la façon suivante. Tout d'abord, un décodage classique de l'ambisonique est effectué sur quatre haut-parleurs. La matrice de décodage "D" définie dans I.4.3.c est appliquée. Elle permet le décodage sur quatre haut-parleurs virtuels utilisant la relation (I.24). Les directions des sources virtuelles sont déterminées par σ_1 et σ_2 et leurs opposées. L'ajout de deux sources virtuelles supplémentaires (4 au total) assure une inversion correcte de la matrice de décodage D, comme détaillé dans (I.24). Enfin, les signaux binauraux sont obtenus par filtrage des signaux des haut-parleurs virtuels avec les HRTF des directions associées à la direction de chacun de ces quatre haut-parleurs.

En complément, afin de prévenir des artefacts audibles à cause des sauts de localisation des sources, les résultats obtenus lors de la phase de localisation sont lissés temporellement et fréquentiellement. Pour éviter le bruit de phase qui peut dégrader la localisation aux hautes fréquences, sa valeur est remplacée de façon progressive par la phase correspondant à un retard pur.

I.6 Évaluation de la qualité perçue des systèmes de reproduction sonore spatialisée

Jusqu'à maintenant, nous avons défini des méthodes permettant un encodage de la scène sonore spatialisée, soit par la prise de son, soit par synthèse. Le but final de ces méthodes étant de pouvoir restituer ces scènes à des auditeurs, il est nécessaire de se munir d'outils permettant leur évaluation d'un point de vue perceptif.

Cette tâche implique de nombreuses dimensions perceptives qui peuvent être ramenées à des attributs propres à cette modalité, tels que (entre autres), le timbre, la localisation, l'immersion.

I.6.1 Principe

L'évaluation de la qualité sonore peut suivre deux approches distinctes. La première est basée sur des mesures objectives du signal sonore, ces mesures étant elles-mêmes basées sur des modèles de la perception. Elles sont obtenues à partir des indicateurs de qualité de restitution et/ou de ressemblance avec un signal cible. L'estimation des paramètres mesurables, tels que la distorsion ou le Rapport Signal-à-Bruit (RSB), permet d'obtenir une première indication. L'utilisation de modèles perceptifs plus ou moins complexes permet l'obtention d'une indication plus adaptée [Beerends and Stemerdink, 1992, Susini et al., 1999].

Une deuxième approche, permettant de prendre réellement en compte la perception sonore, est basée sur l'évaluation subjective. Les évaluations de ce type doivent être établies suivant une méthodologie en trois temps [Lawless and Heymann, 1998, Lawless and Heymann, 1998, Le Bagousse, 2014] :

- définition les paramètres du test,
- mise en place du test,
- analyse et interprétation des résultats.

Les paragraphes qui suivent se focalisent sur cette deuxième approche.

I.6.1.a Comparaison par paires

La comparaison par paires permet d'évaluer directement les dissemblances entre deux stimuli audio que l'auditeur classifie selon une échelle continue, et dont les bornes vont de "très semblables" à "très dissemblables". Ce type d'approche a été utilisé dans diverses applications telles que la caractérisation du timbre instrumental [Grey, 1977] ou la caractérisation spatiale de salles de concert [Blauert and Lindemann, 1986]. Cette méthode comporte l'inconvénient d'être très lourde à mettre en place, car le nombre de présentations pour N échantillons est de $\frac{N(N-1)}{2}$ par sujet, ce qui se traduit par une durée importante de test, et par conséquent, par une fatigue accrue des testeurs, ce qui peut perturber leur attention [Scavone et al., 2001]. La norme AES20-1996 [AES, 1996] détermine que la durée de chaque session ne doit pas dépasser 30 minutes, limitant ainsi à 20 le nombre de stimuli d'environ 20 s chacun.

I.6.1.b Méthodes standardisées

Dans le milieu industriel, il est nécessaire de se référer à des méthodes normalisées. Aujourd'hui, l'Union Internationale des Télécomunications (UIT) dispose de recommandations de mise en place de tests d'écoute en fonction des stimuli et des dispositifs de restitution. Une méthode normalisée utilisant l'analyse par paires est proposée par l'UIT dans la recommandation UIT-R BS.1116 [IUT, 1997]. La recommandation BS.1534-1 [ITU, 2003b] communément connue sous le nom de *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) permet d'adapter la méthode à l'évaluation d'un plus grand nombre de stimuli en simultané. Par ailleurs, la recommandation BS.1284-1 [ITU, 2003a] est une proposition d'harmonisation des tests suivant la recommandation BS.1116-1, dans le but d'une comparaison des résultats issus de différents laboratoires. D'autres propositions, comme celle de l'Union Européenne de Radio-Télévision (EBU)[EBU-tech, 1997], font aussi partie des méthodes normalisées pour l'évaluation de la qualité sonore perçue d'un contenu musical en utilisant une évaluation multi-critères.

UIT-R BS.1116 La recommandation UIT-R BS.1116, intitulée "Méthodes d'évaluation subjective des dégradations faibles dans les systèmes audio y compris les systèmes sonores multivoies", est destinée à l'évaluation des dégradations presque imperceptibles et demande la compétence d'un panel de sujets "experts" afin de déceler ces variations.

Cette recommandation est prévue pour être utilisée lors de l'évaluation de plusieurs attributs de qualité. L'évaluation de chaque variable doit être effectuée de façon indépendante, afin de focaliser l'attention des sujets sur chacune d'entre elles. Elle dépend du dispositif de restitution (tableau I.3). Le document prévoit la notation de la *"qualité audio de base"*, quel que soit le dispositif de restitution, la *"qualité de l'image stéréophonique"*, dans le cas d'une diffusion stéréo et la *"qualité frontale de l'image"* et l'*"impression de qualité*

ambiophonique", dans le cas de l'utilisation d'un système multivoies.

Variable de qualité	Système de restitution		
	monophonique	stéréo	multivoie
Qualité audio de base	*	*	*
Qualité de l'image stéréophonique		*	*
Qualité frontale de l'image			*
Impression de qualité ambiophonique			*

Tableau I.3 : Attributs de qualité pouvant être évalués sous la recommandation UIT-R BS.1116 [IUT, 1997] en fonction du système de reproduction.

La *"qualité frontale de l'image"* et l'*"impression de qualité ambiophonique"* permettent respectivement d'évaluer la localisation de l'image frontale et l'impression d'espace. Ces deux critères sont très peu utilisés, car il y a un manque de connaissances sur leur définition, comme cela est évoqué dans la mise en garde consignée dans le texte de la recommandation.

La méthode utilisée pour l'évaluation d'échantillons dans la recommandation UIT-R BS.1116 est celle du "doublement aveugle à triple stimulus et référence dissimulée", où la référence est connue et présentée comme "A" et où stimulus "B" et "C" sont comparés à cette référence sur une échelle comportant ou ne comportant pas de graduations. Dans ce dernier cas, les résultats doivent être normalisés afin de constituer une échelle comparative des résultats des différents sujets. La normalisation est effectuée selon l'équation

$$Z_i = \frac{(x_i - \bar{x}_i)}{\sigma_i} \cdot \sigma_s + \bar{x}_s \,, \tag{I.34}$$

où Z_i est le résultat normalisé, x_i est la note établie par le participant i, $\bar{x}_i$ est la note moyenne du participant i, $\bar{x}_s$ est la note moyenne de l'ensemble de sujets, σ_i l'écart-type du sujet i et σ_s est l'écart-type global.

Il est à noter que les échantillons d'une durée maximale de 20 s peuvent être écoutés autant de fois que nécessaire avant de déterminer une note, l'utilisateur définissant le rythme de passage des échantillons. Le test doit alors être conçu afin de ne pas dépasser 30 minutes par session, conformément à la norme [AES, 1996].

Les caractéristiques électroacoustiques des systèmes de restitution, leur disposition et le niveau d'écoute sont également spécifiés par cette recommandation. Elle détermine aussi les caractéristiques des locaux en termes de taille, géométrie, ainsi que leurs caractéristiques acoustiques.

BS.1534-1 La recommandation BS.1534-1 ou MUSHRA intitulée "Méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes de codage" est basée sur la recommandation UIT-R BS.1116 et peut être utilisée pour l'évaluation des critères évoqués dans cette dernière. Comme son nom l'indique, cette méthode a été conçue pour l'évaluation des dégradations importantes [Soulodre and Lavoie, 1999] engendrées notamment par l'encodage et la transmission du signal, telles que la limitation de bande, le rajout de bruit, les pertes du signal et la perte de paquets. Elle prévoit également l'étude de faibles dégradations à condition d'utiliser un panel de sujets experts.

La méthode MUSHRA est un test en double aveugle à stimuli multiples, avec une référence cachée, et un ou plusieurs repères cachés (ancres ou signaux dont la qualité est connue). A la différence de la recommandation BS.1116-1 où deux signaux sont comparés entre eux, la méthode MUSHRA permet de comparer plusieurs stimuli simultanément (15 au maximum). L'utilisation d'une référence cachée joue ici un double rôle. D'une part, elle permet de fixer le haut de l'échelle, et d'autre part, elle permet de contrôler la pertinence des réponses lorsque les dégradations sont faibles. Les signaux dégradés sont comparés à cette référence et entre eux, donnant ainsi une notation relative à l'ensemble de signaux comparés.

L'évaluation se fait sur une échelle issue de la recommandation UIT-R BT.500 [IUT, 2012]. Il s'agit d'une échelle continue à cinq valeurs allant de mauvais à excellent (figure I.18).

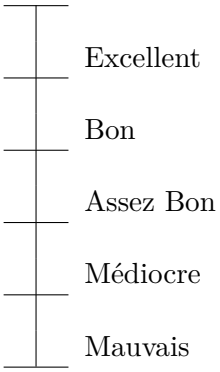


Figure I.18 : Échelle d'évaluation utilisée pour la méthode MUSHRA issue de la recommandation UIT-R BT.500 [IUT, 2012].

Vers une adaptation des normes aux contenus spatialisés Conscient que les méthodologies proposées par les recommandations de l'UIT ne satisfont pas pleinement les contraintes liées à la perception du son spatialisé, Le Bagousse et al [Le Bagousse et al., 2010] revisitent les normes pour les adapter à une évaluation plus précise des contenus audio spatialisés. Trois dimensions principales ont été extraites, permettant d'évaluer la qualité d'un contenu sonore spatialisé grâce à une étude sémantique multidimensionnelle. Le Bagousse [Le Bagousse and Paquier M., 2011, Le Bagousse et al., 2012, Le Bagousse, 2014] a mis en place une méthode inspirée principalement de la recommandation UIT-R BT.500

MUSHRA, en utilisant les attributs issus de l'étude [Le Bagousse et al., 2010] (ie. Qualité globale, timbre, espace et défauts), ceci afin d'évaluer les scènes sonores spatialisées pour une représentation sur un système 5.1 sur haut-parleurs et binauralisé sur casque.

Dans la méthodologie proposée par Le Bagousse, la notation est faite sur une échelle non graduée. En effet, il a été démontré à plusieurs reprises [Zielinski et al., 2008, Zielinski et al., 2007, Jones and McManus, 1986] que la graduation des échelles introduit des biais. A titre d'exemple, la valeur sémantique des adjectifs utilisés n'est pas équidistante et leur valeur change en fonction de la langue utilisée. Afin de fournir de repères suffisants aux sujets, la méthode proposée par Guski [Guski, 1997] est adoptée par Le Bagousse. Des étiquettes sont affichées uniquement aux extrémités de l'échelle, nommées "haute qualité" et "basse qualité".

La recommandation UIT-R BT.500 MUSHRA préconise une ancre de basse qualité permettant de définir la dynamique de la plage d'évaluation. A ce propos, Le Bagousse introduit une ancre spécifique qui cherche à estimer uniquement l'attribut à évaluer.

Dans une première étude, l'auteur propose les ancrages suivants :

Ancre	**5.1**	**Binaural**
Timbre	Filtrage à 3.5 kHz	
Défauts	Ajout de bruit rose	Ajout de bruit rose et clics
Espace	Inversion canal avant droit R et arrière gauche Ls	Inversion des canaux R et L par portion + passages mono

Tableau I.4 : Ancrages spécifiques à chaque attribut proposés par Le Bagousse lors d'une évaluation d'écoute sur un système 5.1 ou en écoute sur casque après la binauralisation des signaux 5.1.

Ces trois ancrages font partie des corpus des signaux test et sont présentés dans les différentes phases d'évaluation. Après analyse, il est remarqué lors de l'écoute sur système 5.1 que l'ajout du bruit rose est le seul ancrage valable de façon indépendante, car les autres ancrages affectent également l'appréciation des autres attributs (figure I.19).

Lors de l'évaluation de la binauralisation du système 5.1, les différents ancrages laissent apparaître la pertinence de leur choix pour l'évaluation des défauts et du timbre. D'autre part, les différentes ancres sont notées au milieu de l'échelle pour l'évaluation de l'espace, démontrant par la même que l'ancrage spatial proposé n'est pas plus pertinent que la réduction monophonique ou que la réduction spatiale simple (figure I.20).

Dans une troisième étude, l'auteur démontre que l'utilisation d'un ancrage unique évaluant l'ensemble des attributs est suffisante et permet d'atteindre des conclusions comparables à celles obtenues avec des ancrages indépendants, simplifiant ainsi le test et réduisant le

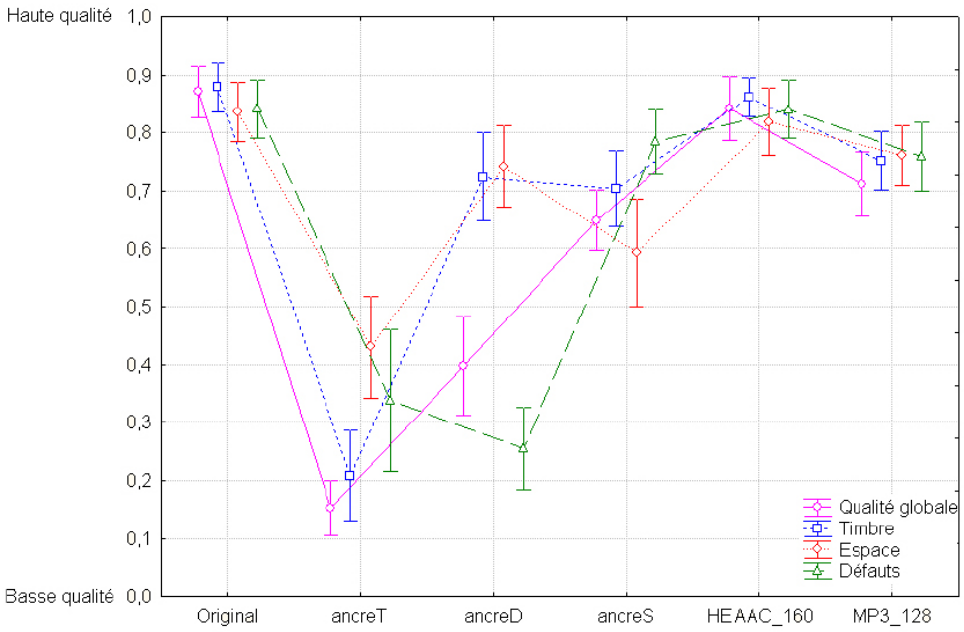


Figure I.19 : Moyennes et intervalles de confiance à 95% des notes obtenues par les différentes versions pour chaque attribut évalué de l'écoute sur système 5.1 utilisant les ancrages définis dans la colonne correspondante sur le tableau I.4. Extrait de [Le Bagousse, 2014].

corpus de signaux, et par conséquence la durée du test.

La méthodologie proposée par Le Bagousse a été largement décrite et nous l'utilisons dans la suite de notre étude (chapitre III). Cette méthode a été choisie parmi celles existantes, car elle permet une mise en place relativement simple prenant en compte des axes perceptifs souvent négligés et s'appuyant sur des méthodes normalisées.

Actuellement, avec l'expansion des technologies de spatialisation sonore, il existe un besoin de méthodes normalisées multicritères permettant d'évaluer la qualité de l'expérience d'écoute d'un contenu sonore spatialisé. Ainsi, le groupe de travail BiLi [Bili, 2014] a récemment fixé une feuille de route sur l'ensemble d'attributs à prendre en compte afin d'établir une méthodologie adaptée à l'évaluation de la qualité du rendu d'une scène sonore spatialisée, et notamment lors d'une restitution binaurale [Nicol et al., 2014].

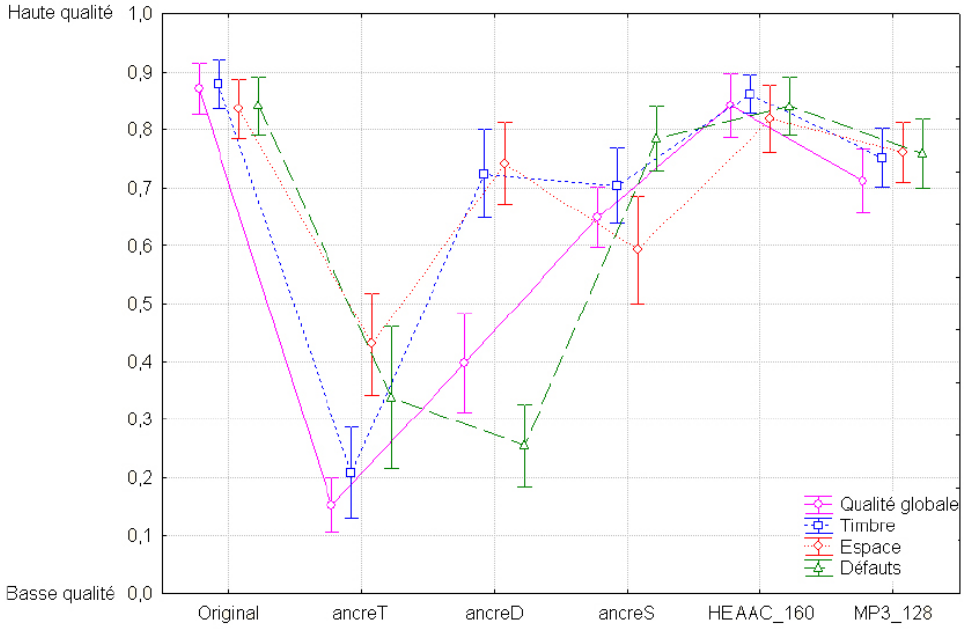


Figure I.20 : Moyennes et intervalles de confiance à 95% des notes obtenues par les différentes versions pour chaque attribut évalué de l'écoute binaurale, utilisant les ancrages définis dans la colonne correspondante sur le tableau I.4. Extrait de [Le Bagousse, 2014].

I.7 Conclusion

Ce chapitre propose un état de l'art des méthodes de spatialisation sonore, ainsi que des formalismes physiques et perceptifs qui rentrent en jeu dans leur mise en œuvre. Nous avons établi un recueil de méthodes de captation et restitution de la scène sonore ainsi que des méthodes de transport ou encodage permettant de faire le lien entre les deux bouts de chaîne.

Ces méthodes ont été classées de deux façons : une première, en fonction du mode dont les informations représentant la scène sonore sont extraites et une deuxième, en fonction de la représentation de la scène sonore.

La première classification nous a permis de déterminer les familles de méthodes : "physiques", "perceptives" et "mixtes" et la deuxième a été également divisée en trois catégories : *channel based*, *soundfield based*, et *object-based*.

Le destinataire final, des procédés ici décrits étant un auditeur, nous avons également effectué une synthèse des méthodes permettant l'évaluation de la qualité perçue du rendu sonore.

Au regard de l'objectif principal de cette thèse, nous pouvons retenir deux pistes essentielles pour la suite : [format de liste]

- d'une part, les avantages que présentent les formats objets dans la représentation d'une scène sonore, à la fois en matière de compacité et de leur indépendance par rapport au dispositif de restitution,
- d'autre part, la versatilité du binaural comme outil universel de restitution sonore,

En effet, nous avons montré comment cette technologie permet de rendre compatible l'ensemble des méthodes de restitution sonore sur un dispositif aussi simple et léger qu'un casque stéréophonique.

Comme il a été démontré dans l'état de l'art, ces deux pistes, ne se réduisent pas à une méthode unique, laissant la porte ouverte à de nouvelles voies à explorer en termes d'amélioration et d'optimisation des traitements. Ces derniers doivent être réalisés dans le but d'une amélioration de la qualité perçue en fin de chaîne acoustique par l'auditeur final.

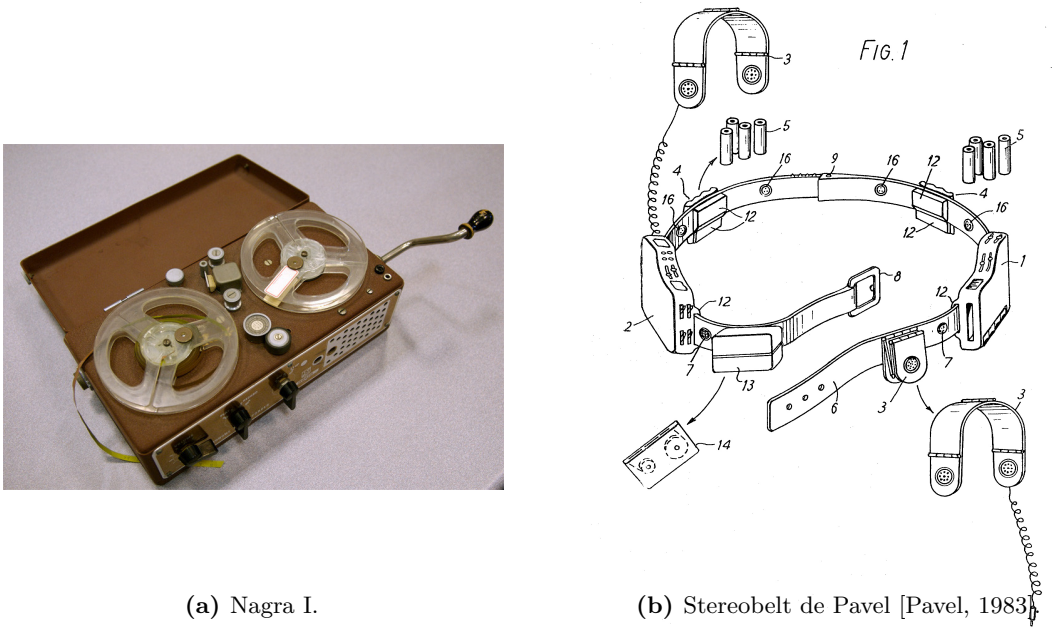
II Le son 3D pour les terminaux mobiles

La problématique principale de cette thèse porte sur l'implémentation de technologies du son spatialisé sur des terminaux mobiles. Avant de décrire les solutions proposées, il est tout d'abord nécessaire de comprendre les spécificités et les contraintes liées à ce genre de dispositifs. En effet, l'ensemble des techniques permettant la captation, le traitement et la restitution du son spatialisé, détaillées en I.3, I.4 et I.5, présentent des contraintes particulières pouvant les rendre incompatibles avec des dispositifs nomades. Une fois ces constats posés, il conviendra de proposer des nouvelles méthodologies pouvant dépasser ces limitations.

II.1 Spécificités et contraintes propres aux terminaux mobiles

Depuis 1951, avec l'invention du premier magnétophone portatif Nagra I par le Suisse Stefan Kudelski [Schoenherr, 2005], la prise de son en mobilité n'a cessé d'évoluer. Malgré les 7 kg [1] que les preneurs de son devaient porter en bandoulière [Nagra, 2014] (figure II.1a), des équipements comme le Nagra III ont contribué au développement de la prise de son en extérieur, ce qui a changé radicalement les méthodes de travail de la radio et du cinéma, en leur offrant des outils techniques pour la naissance de courants comme "la Nouvelle Vague", ce qui a valut à son créateur la reconnaissance internationale par des Oscars en 1965, 1977, 1978 et 1990, et des Grammy en 1984 et 1986. Dans le domaine de la restitution audio en mobilité, l'entreprise Japonaise Sony s'est emparé du marché grand public dès les années 80, grâce au radio cassette portatif Walkman, d'après le brevet de Pavel [Pavel, 1983], imposant l'utilisation du casque comme l'accessoire indispensable

[1] Sans compter le poids de batteries de rechange, microphones, câbles et les bobines de bande magnétique.



(a) Nagra I. **(b)** Stereobelt de Pavel [Pavel, 1983]

Figure II.1 : Premiers dispositifs portatifs de captation et de restitution sonore.

pour l'écoute de dispositifs portatifs et devenant immédiatement un accessoire de mode [Committees, 2000]. La famille de dispositifs portatifs Walkman comporte plus de 500 modèles et a su s'adapter aux avancées technologiques, proposant ainsi en 1988 le premier lecteur de disques compacts, le DiscMan D-88 en 1991, le lecteur enregistreur portatif à cassette audio numérique *Digital Audio Tape* (DAT) Walkman TCD-D3 en 1990 et en 1992 le MiniDisc [Schoenherr, 2005, Sanderson and Uzumeri, 1997, du Gay et al., 2013]. L'engouement du grand public pour les supports dématérialisés, tels que le MP3, se généralise avec l'apparition de l'iPod de Apple en 2001 et le lancement de la plateforme de téléchargement iTunes en 2003. L'utilisation des téléphones portables comme des dispositifs de reproduction de fichiers multimédia ne s'est pas faite attendre. Blackberry a ainsi conquis les marchés de masse avec ses dispositifs (8800, Pearl, Pearl Flip et Curve), se plaçant en tête du marché aux Amériques (du Sud, Centrale et du Nord) avec 45% de parts de marché jusqu'à l'apparition de l'iPhone en 2007.

Aujourd'hui, les *smartphones* sont devenus de véritables micro-ordinateurs portatifs avec des capacités comparables à celles des ordinateurs de bureau quelques années auparavant. La puissance de calcul ayant un impact sur la consommation énergétique et par conséquence sur l'autonomie, ce facteur continue à être limitant et impose des simplifications des méthodes algorithmiques utilisées.

Avec l'introduction des *smartphones* et la connectivité de ce type de dispositifs, c'est une nouvelle façon de "consommer" des contenus radiophoniques et télévisuels qui a vu le jour. Une enquête menée par Vogel en 2008 [Vogel et al., 2008] a révélé que la quasi-totalité de la population mondiale entre 12 et 18 ans possède et utilise un lecteur multimédia portatif et B. Gutiérrez Camarero et al. [Gutiérrez Camarero and Moledero Dominguez, 2007]

ont montré que 80% de la population des 12-30 ans écoutent quotidiennement des contenus multimédia avec des casques ou avec des oreillettes. Le principal distributeur de musique dématérialisé, iTunes, annonçait déjà en 2007 que 2 millions de long-métrages avaient été téléchargés [Wolfson and Neumayr, 2008], et, dès la fin de l'année 2008, le chiffre de 200 millions d'épisodes de téléfilms distribués a été mis en avant par ce même distributeur. En février 2013, le géant du téléchargement a annoncé que la barre de 25 milliards de morceaux vendus sur leur plateforme avait été atteinte [Monaghan and Garlinghouse, 2013].

II.1.1 L'audio dans les terminaux mobiles

Pour des applications audio numériques, notamment pour la Musique Assistée par Ordinateur (MAO), les *smartphones* et autres terminaux mobiles sont devenus de véritables instruments de musique, grâce à des applications dédiées ou comme interface homme-machine. Aujourd'hui, ils permettent le contrôle de dispositifs audio utilisant des protocoles de communication tels que le *Music Instrument Digital Interface* (MIDI) et le *Open Sound Control* (OSC).

Malgré l'évolution des capacités des terminaux mobiles, les dispositifs de captation et restitution sonore intégrés dans les téléphones portables restent majoritairement monophoniques, sauf quelques rares exceptions.

II.1.1.a Prise de son

Paradoxalement, les téléphones portables disposent de **multiples microphones**. Aujourd'hui, ces transducteurs sont fabriqués généralement avec une technologie *Micro-electro-mechanical Systems* (MEMS). Cette technologie permet de réaliser des microphones de très petite taille, en grande série, comportant des caractéristiques relativement homogènes et à faible coût (Annexe B). Malheureusement, la façon dont ces microphones sont implémentés sur les terminaux modifie leurs propriétés acoustiques (directivité et réponse en fréquence), les rendant difficilement utilisables pour des applications audio avancées. Les microphones et leur emplacement ont en effet été principalement conçus pour la captation sonore de la voix lors d'une communication téléphonique. Ils sont utilisés simultanément ou alternativement, en fonction du mode de communication (mode "auriculaire [2]", ou mode "haut-parleur[3]") et sont couplés à des algorithmes de beamforming, débruitage et/ou d'annulation d'écho [Dahl and Claesson, 1999]. Généralement, ces algorithmes sont intégrés directement dans la couche matérielle [Byun, 2009] et les signaux microphoniques ne sont pas accessibles depuis les couches logicielles des systèmes d'exploitation des terminaux [Haidar, 2012].

[2] Téléphone tenu à la main, haut-parleur près de l'oreille et microphone à proximité de la bouche.

[3] Téléphone éloigné, tenu à la main, ou posé sur une surface, utilisant les transducteurs insérés dans le dispositif. A ne pas confondre avec le mode "main libres" qui nécessite un accessoire comportant lui-même des transducteurs.

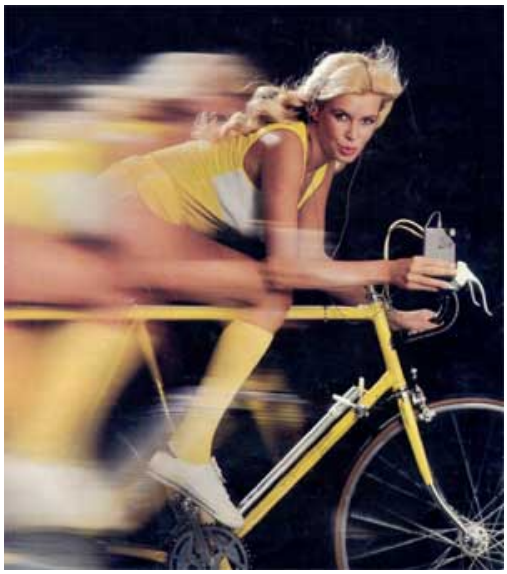
Ce choix est fait de la part de constructeurs, car les applications audio sont très gourmandes en CPU [Lu et al., 2010]. Le décodage et la lecture d'un fichier audio encodé en AAC à 256 kbps utilisent en moyenne 8% du CPU d'un iPhone 4. Sur ce même terminal, des tâches de classification avec des modèles gaussiens atteignent des pourcentages d'utilisation du processeur allant entre 22 et 36%, en utilisant des signaux audio échantillonnés à 8 kHz uniquement. L'utilisation de couches matérielles dédiées au traitement de l'audio en entrée et sortie du terminal permet de décharger le CPU de ces tâches, améliorant les performances et réduisant considérablement la consommation énergétique [Paulin et al., 1997].

Certains dispositifs disposent d'entrées audio analogiques (connecteur du kit "mains libres") et numériques (dock Apple)[Lin and Sung, 2009], généralement limitées à un ou deux canaux. On trouve également des systèmes d'interfaçage permettant l'utilisation de cartes son externes (connecteur OTG [Remple, 2003] et le *camera connection kit* d'Apple). Ces solutions rendent possible une acquisition d'un nombre plus important de signaux avec une qualité contrôlée. En effet, les signaux ainsi captés profitent des CAN et des préamplificateurs de la carte son, aux dépens de la portabilité des dispositifs. L'utilisation des périphériques de ce type nécessite une connectique adaptée et généralement une source électrique complémentaire, car le terminal mobile ne délivre pas la puissance électrique suffisante.

II.1.1.b Restitution du son

En matière de **restitution sonore**, les *smartphones* sont généralement munis de deux haut-parleurs, l'un permettant la communication "auriculaire" et le deuxième utilisé dans le mode "haut-parleur" ou pour la reproduction de contenus multimédia. Certaines tablettes tactiles proposent une diffusion sonore améliorée utilisant des dispositifs multi-haut-parleurs disposés sur leur cadre et dont l'utilisation est pilotée directement par la couche matérielle [press release, 2011]. Le manque de place dans ces dispositifs fait que la reproduction sonore n'est effectuée que sur une bande passante réduite. La technologie des MEMS est une piste prometteuse pour la diffusion sonore sur ce genre de dispositifs, faisant des haut-parleurs numériques une réalité [Cohen et al., 2010].

De la même manière que pour l'acquisition, les terminaux mobiles sont munis de **sorties audio analogiques et numériques**. Tous les dispositifs mobiles sont en particulier équipés de prises stéréo analogiques pouvant être utilisées pour la diffusion sur un terminal HiFi adapté. Des interfaces sans fil "Bluetooth" [Bluetooth Audio Video Working Group, 2002] permettent également la transmission de signaux stéréophoniques utilisant un protocole de communication et un encodage spécifique. Aujourd'hui, une large gamme de dispositifs permet la diffusion des contenus stéréophoniques des produits Apple en utilisant les signaux analogiques ou numériques de ces terminaux, et des adaptateurs HDMI permettent la lecture jusqu'à huit canaux audio sans compression.



(a) Walkman de Sony au début des années 80.

(b) Beats avec Dr Dre comme image de marque 2014.

Figure II.2 : Images publicitaires du casque comme accessoire de mode.

La prise stéréophonique des terminaux mobiles a été principalement conçue pour l'utilisation d'un casque, car ce dernier est le dispositif de restitution privilégié pour ce genre de dispositifs. A noter que depuis l'apparition du walkman de Sony, le casque est devenu un accessoire de mode [Committees, 2000] évoquant le sport et les loisirs [du Gay et al., 2013] et est devenu un symbole social [Joseph, 1980] utilisant l'image de personnalités publiques pour atteindre une large population (figure II.2).

Aujourd'hui, la large utilisation de téléphones portables, *smartphones* et lecteurs multimédia portatifs a favorisé l'utilisation du casque comme le dispositif le mieux adapté à la reproduction sonore en mobilité. Avec l'évolution de la compression audio et la miniaturisation de l'électronique, les utilisateurs peuvent transporter l'ensemble de leur médiathèque ou accéder continuellement à des contenus disponibles sur la toile.

L'annonce de l'achat de la société *Beats electronics* le 14 mai dernier (2014) par Apple, pour 3.2 milliards de dollars [Smith et al., 2014], ne fait que confirmer l'importance du casque pour la consommation de la musique, et montre que l'utilisation de celui-ci n'est plus un frein pour l'introduction du binaural dans le marché grand public.

II.2 Le son 3D pour les terminaux mobiles : une contrainte de taille

Dans le chapitre I, nous avons établi un recueil des technologies et méthodes permettant la captation, le traitement et la restitution des scènes sonores spatialisées. En I.5, nous avons détaillé les techniques ayant atteint une maturité les rendant directement disponibles pour être implémentées, voire déjà présentes sur le marché. Nous souhaitons maintenant pouvoir appliquer les techniques de captation et de restitution d'une scène sonore spatialisée aux contraintes et aux limitations des terminaux mobiles.

II.2.1 Prise de son 3D

Les techniques de captation spatialisée nécessitent des dispositifs spécifiques afin de capter l'ensemble des informations spatiales de la scène sonore.

Les technologies stéréophoniques coïncidentes ont déjà fait leurs preuves dans la captation en mobilité. Cependant, elles offrent une restitution réduite de la scène sonore. Les autres méthodes stéréophoniques et multicanales utilisent des dispositifs très encombrants les rendant incompatibles avec les terminaux mobiles.

Le binaural peut être une solution envisageable avec l'utilisation de microphones à poser au niveau des oreilles du preneur de son. Comme indiqué en I.4.2, l'utilisation de cette technique implique des contraintes et génère d'importants inconvénients. Les autres solutions de captation du binaural étant très encombrantes, elles ne sont pas adaptées à la prise sonore en mobilité, même si elles n'utilisent que deux canaux sonores.

Les méthodes ambisoniques utilisent des dispositifs relativement compacts (ie. *soundfield*, *eigenmike*), comme décrit en I.4.3.b. Se pose alors le problème du nombre de signaux nécessaires à la description d'une scène sonore. A l'ordre 1, les quatre signaux de l'ambisonique peuvent éventuellement convenir à l'utilisation de ce type de dispositifs, à condition d'utiliser une interface et/ou une connectique adaptée. Aux ordres supérieurs, le grand nombre de signaux nécessaires est une limitation importante. Les quantités élevées de données à manipuler les rendent incompatibles avec ce type de dispositifs, en matière de CPU, de capacité de stockage et de transmission des données [Paulin et al., 1997]. Cependant, l'ambisonique étant compatible avec le binaural, grâce aux méthodes de décodage (I.4.3.c), et notamment aux méthodes de décodage actif (I.5.2.b), l'ambisonique est une piste prometteuse pour leur utilisation en mobilité.

II.2.2 Restitution sonore 3D

La restitution sur haut-parleur est difficile au regard des contraintes imposées par les terminaux mobiles, et notamment leur encombrement. La taille des haut-parleurs et enceintes acoustiques étant directement proportionnelle à leur bande passante et leur efficacité acoustique, la qualité du rendu sonore est limitée lorsque les haut-parleurs sont de faible taille. Cette limitation importante nous impose d'écarter ce choix de reproduction. Qui plus est, les systèmes multi-haut-parleurs imposent des contraintes sur le positionnement des transducteurs et sur le lieu d'écoute, contraintes qui vont à l'encontre de l'utilisation en mobilité.

Tel qu'il a été décrit en II.1.1.b, le casque stéréophonique est le mode de restitution privilégié pour les dispositifs mobiles, de par sa taille et les avantages qu'il offre en termes d'intimité et confidentialité. Le binaural se servant du casque stéréophonique pour sa restitution est par conséquent la méthode de reproduction à privilégier pour la restitution sonore spatialisée en mobilité (I.4.2).

Comme précisé en I.5.2, le binaural permet par ailleurs le rendu direct de contenus binauraux et de toute technique effectuant une restitution sur haut-parleurs moyennant un traitement spécifique.

Il est possible de trouver sur le marché des applications se servant d'un moteur de restitution binaurale comme le jeu vidéo (100% audio et sans vidéo) pour iOS *Papa Sangre* de la société *Somethin'Else*. Cette application utilise le moteur *Papaengine* pour activer les événements sonores qui sont "binauralisés" de façon dynamique au moment de la restitution. Il est piloté grâce aux capteurs de localisation du dispositif mobile. Ces modalités d'écoute font également une incursion dans le milieu culturel et patrimonial lors des visites commentées à réalité augmentée. Le système *Chimera* de "l'Agence du Verbe" [Rueff, 2010] permettant ainsi la reproduction de contenus binauraux pré-enregistrés à partir des données GPS, en suivant des itinéraires particuliers au gré des envies des visiteurs.

Des moteurs de "binauralisation" directement implémentés sur les lecteurs multimédia permettent le rendu des contenus 5.1 et 7.1 sur un casque stéréo avec une utilisation transparente pour l'utilisateur [radiofrance.fr, 2014]. Le moteur de "binauralisation" peut également être implémenté au niveau du casque. Dans ce cas particulier, il faut cependant assurer le transit de l'ensemble du flux audio vers le casque. L'utilisation des connections par USB ou HDMI permet de récupérer le flux numérique, et le décodage, la binauralisation, ainsi que la conversion analogique-numérique, doivent s'effectuer au niveau du casque.

Des moteurs de "binauralisation" de contenus ambisoniques et WFS peuvent également être implémentés dans les terminaux. Ils doivent être développés de façon simplifiée afin

de ne pas surcharger la CPU des dispositifs (I.5.2).

II.2.3 Formats audio 3D

Le binaural présente l'avantage d'effectuer l'encodage de la scène sonore complète en se servant uniquement de deux signaux audio. Cette solution est alléchante, mais elle offre une image fixe et dépendante de la morphologie ou des HRTF qui ont été utilisées à sa création.

La "binauralisation" des contenus audio spatialisés destinés au rendu sur haut-parleurs est possible grâce à la technique des sources virtuelles. Elle permet un rendu de bonne qualité, à condition d'utiliser une méthode efficace et des HRTF adaptées à la morphologie de l'auditeur.

La généralisation de cette méthode demande un format universel pour le stockage des HRTF individuelles et des signaux audio. En effet, lorsqu'il s'agit d'un faible nombre de haut-parleurs virtuels (stéréo, 5.1, 7.1) ou de l'ambisonique à l'ordre 1, les méthodes d'encodage traditionnelles *Channel-based* décrites en I.5.1.a peuvent suffire pour une utilisation sur un terminal mobile. Lors de l'utilisation de systèmes plus complexes (10.2, 22.2 et HOA), une description paramétrique de la scène sonore (format objet par exemple décrits en I.5.1.c) sera cependant plus adaptée aux besoins et aux contraintes de ce type de terminaux.

II.3 Première esquisse de chaîne audio 3D pour les terminaux mobiles

Prenant en compte d'une part, les outils du son 3D disponibles aujourd'hui, et d'autre part, les contraintes des terminaux mobiles, nous avons souhaité réaliser une première maquette permettant de démontrer la faisabilité de la captation et de la restitution sonore sur un terminal mobile, mettant en parallèle différentes méthodes de spatialisation sonore.

Cette maquette est décrite dans le chapitre III. Elle offre la possibilité d'évaluer sa qualité sonore perçue, nous permettant de la comparer avec des différentes méthodes de spatialisation sonore pouvant être utilisées dans un terminal mobile.

Pour la prise de son, un microphone *soundfield* est utilisé et un décodage des signaux ambisoniques à l'ordre 1 permet leur rendu binaural sur casque stéréophonique. Afin d'évaluer la qualité de cette maquette, trois prises de son (stéréophonique, binaurale et ambisonique à l'ordre 1) ont été effectuées. Pour la reproduction de l'ambisonique, deux décodages ont été implémentés.

Une première analyse nous permettra par la suite d'identifier que le maillon faible de la chaîne sonore est le dispositif de prise de son et son décodage. Ce constat nous conduira à proposer une nouvelle méthodologie de captation, plus adaptée aux terminaux mobiles (chapitres IV et V).

Figure II.3 : Dispositif de démonstration de la diffusion en direct sur tablette tactile avec restitution binaurale : lors de l'événement "l'Opéra sur tous les écrans", au printemps 2011 à Rennes [Guizart, 2011]

III Évaluation d'une première maquette de chaîne de reproduction sonore 3D pour les terminaux mobiles

Le constat effectué au chapitre II nous a permis de définir une première maquette, représentant une chaîne de captation et de restitution du son spatialisé sur des terminaux mobiles. Dans ce chapitre, nous procédons à son évaluation subjective et objective. A cette fin, des événements sonores ont été enregistrés à l'aide de cette maquette comportant un microphone ambisonique à l'ordre 1 et une "binauralisation" ou décodage pour une restitution sur casque.

Afin d'évaluer la maquette en question, en la comparant avec d'autres techniques, une base de données d'enregistrements a été constituée à l'aide de trois dispositifs de captation : une tête acoustique, un microphone *Soundfield* et une paire stéréo AB.

Les signaux ont été testés sur casque. L'utilisation d'une tête acoustique permet de les comparer avec une "binauralisation" naturelle et directe, qui représente *a priori* une version de référence.

La paire stéréo peut être vue comme une référence culturelle, la plupart des contenus disponibles et écoutés au casque étant enregistrés aujourd'hui avec des techniques stéréophoniques.

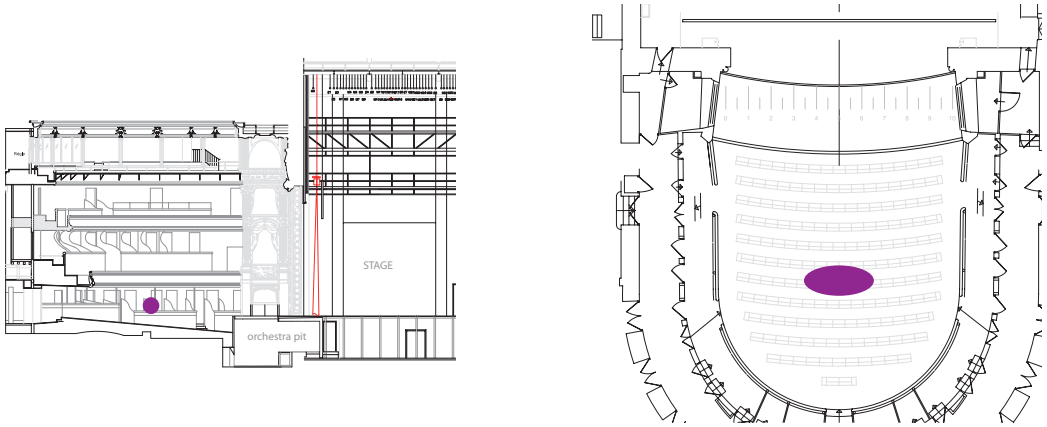
Afin de constituer le corpus d'échantillons sonores utilisés lors du test, nous avons saisi l'opportunité de participer à l'événement "l'Opéra sur tous les écrans", au printemps 2011 à Rennes [Guizart, 2011], événement qui coïncidait avec le début et contexte de cette thèse.

En effet, l'objectif de cet événement était de proposer des méthodes alternatives per-

mettant d'assister à des spectacles vivants de façon virtuelle et interactive. Parmi les différentes méthodes de transmission et de diffusion en direct qui ont été mises en œuvre, une transmission en direct a été réalisée sur des tablettes tactiles avec rendu audio sur casque comme illustré en figure II.3. Dans sa mise en œuvre, l'utilisateur était libre de choisir le point de vue, parmi plusieurs qui lui étaient proposés, l'audio assurant l'expérience immersive. Dans cette optique, la prise de son binaurale avec la tête acoustique a été retenue pour la transmission sonore en direct, car cette méthode présentait *a priori* la meilleure qualité pour l'objectif recherché.

En parallèle, une régie radio et télévision a permis la transmission traditionnelle sur la radio nationale (Radio France), sur la télévision régionale (France 3 Bretagne), sur la chaîne nationale (Mezzo) et sur écran géant, place de l'Opéra à Rennes.

Les résultats exposés dans ce chapitre ont fait objet d'une présentation à la conférence *Acoustics 2012* et sont consignés dans [Palacino et al., 2012].



(a) Opéra coupe longitudinale. **(b)** Plan d'orchestre.

Figure III.1 : Position des dispositifs de prise de son au centre du parterre-public de l'opéra de Rennes (surface mauve).

III.1 Dispositif expérimental

Afin de se rapprocher au mieux des conditions réelles d'écoute de l'opéra, les transducteurs utilisés ont été disposés dans la salle, au milieu du public. Comme illustré en figure III.1, ils ont été placés au plus près les uns des autres, afin de capter un champ de pression acoustique quasi-identique, à une hauteur de 1,3 m (conditions d'écoute d'un spectateur assis). L'emplacement illustré en figure III.1 se trouve au parterre à 8,5 m de la scène et à 5,5 m de la fosse d'orchestre. Deux fauteuils ont été réservés de part et d'autre afin de réduire les bruit parasites engendrés par les spectateurs et afin de protéger le dispositif (figure III.2).

III.1.1 Systèmes de prise de son

Les enregistrement binauraux ont été effectués avec une tête acoustique KU100 de Neumann, l'enregistrement ambisonique avec un microphone Soundfield équipé du décodeur permettant d'obtenir directement les signaux au format B. La paire stéréophonique était composée de deux microphones omnidirectionnels 103V 4003 DPA écartés de 0,30 m (figure III.2). Tous les signaux ont été conditionnés à l'aide de pré-amplificateurs audio adaptés. Les signaux ont été numérisés à une fréquence d'échantillonnage de 48 kHz et avec une résolution de 24 bits, avec un CAN RME Optamic II.

(a) **(b)**

(c) **(d)**

Figure III.2 : (a), (b), (c) Dispositif de prise de son et (d) régie technique.

Les signaux numérisés étaient transmis par voie optique de la salle à la régie. En effet, le CAN comporte une sortie optique au format ADAT. Ce protocole permet la transmission jusqu'à 8 signaux audio numérisés à 48 kHz et codés sur 24 bits sans aucune compression, ainsi que la synchronisation des horloges des différentes machines raccordées à l'émetteur qui fait office d'horloge maître. La transmission, limitée à 10 m, se fait par fibre optique munie de fiches Tostlink utilisant une lumière dans le spectre visible. Pour nous affranchir de la limitation de distance imposée par le protocole, nous avons dû utiliser un dispositif développé à cet effet : les signaux ADAT ont été transformés en signaux électriques avec l'utilisation de transducteurs opto-éléctriques Firecomms F01202. Chaque signal ainsi obtenu a été de nouveau converti en un signal optique de 1410 nm à l'aide d'un transducteur électro-optique. Ce signal était transmis par un réseau en fibre optique mono-mode. Enfin, la conversion inverse était réalisée à la régie afin d'obtenir le signal en lumière rouge compatible avec le protocole ADAT.

L'ensemble des signaux microphoniques transitant par le réseau optique a été stocké. Les signaux binauraux, en plus d'être enregistrés, ont été également transmis à la régie d'intégration pour la diffusion en direct sur des tablettes aux différents points de la ville de Rennes.

III.1.2 Décodage binaural du microphone ambisonique

L'objet de la maquette ici présentée porte principalement sur la restitution binaurale de l'ambisonique. Pour rendre possible cette restitution, il est nécessaire d'avoir recours aux méthodes de décodage adaptées. Les différentes techniques de décodage utilisant des méthodologies qui leur sont propres n'aboutissent pas aux mêmes résultats en termes perceptifs et du signal. A cet objet, deux méthodes de décodage ont été mises en place afin de pouvoir les évaluer.

La première effectue une projection des HRTF sur la base d'harmoniques sphériques (pour plus de détails consulter I.4.3.c) et est identifiée par la suite comme *"SF dec 1"*.

La deuxième méthode est basée sur un décodage actif de l'ambisonique. Elle effectue une localisation des sources dans la première phase, pour ensuite réaliser un décodage classique sur les haut-parleurs virtuels correspondant aux directions obtenues lors de la première analyse. Ce procédé décrit en I.5.2.b est identifié par la suite comme *"SF dec 2"*.

Les deux décodages ont été implémentés à l'aide de plugins *VirtualStudioTechnology* (VST) dans une interface *Reaper*. Le premier a été développé à Orange Labs et utilise la base de HRTF Pernaux [Pernaux, 2003]. Le second a été mis en place en utilisant l'outil *Harpex* [Berge and Barrett, 2014] de Berge utilisant la base de HRTF "Listen" [IRCAM, 2014] de l'IRCAM.

III.1.3 Postproduction

Afin de retranscrire au mieux la sensation auditive d'un auditeur dans la salle, des tests ont été effectués pendant les répétitions, afin de définir les gains à apporter aux préamplificateurs pour profiter au maximun de la dynamique offerte par le dispositif expérimental. Finalement, un compresseur a été utilisé en fin de chaîne afin d'éviter toute saturation induite lors des phases d'applaudissements du public. En effet, le niveau sonore enregistré pendant ces phases s'est avéré 10 dB plus élevé que le niveau maximal lié à la musique.

La diffusion de l'audio sur les tablettes tactiles se faisant sur casque, un post-traitement des signaux sonores décrits en III.1.2 a été nécessaire pour obtenir deux paires de signaux binauraux (identifiés par la suite par les labels *"SF dec 1"* et *"SF dec 2"*). Enfin, pour la diffusion en direct, les signaux binauraux issus de la tête acoustique ont été légèrement

égalisés selon un protocole mis au point par un ingénieur du son spécialisé dans ce type d'enregistrements. Les signaux issus de la paire stéréo ont été stockés sans aucun post-traitement.

Parallèlement à l'enregistrement des signaux comportant les post-traitements décrits précédemment, les signaux bruts (sans post-traitement) ont été enregistrés pour une utilisation ultérieure.

III.2 Test d'écoute

L'objectif principal de cette expérience est de comparer différentes méthodes de captation d'une scène sonore 3D, dans le but d'une restitution sur casque. Les enregistrements effectués dans le cadre de la prise de son de l'opéra de Mozart *L'enlèvement au sérail* nous ont permis de constituer un corpus de signaux audio composé de cinq "versions" différentes. Elles correspondent aux trois techniques de captation (ambisonique, binaurale et stéréophonique), l'ambisonique étant associé à deux décodages différents et à un ancrage de basse qualité.

Dans un premier temps, un test a été mené, afin de déterminer la qualité du rendu sur casque prenant en compte la qualité globale ainsi que le caractère spatial et timbral perçu. Les résultats ont été ensuite analysés pour déterminer quelle technologie est la plus adaptée pour l'enregistrement d'une scène sonore en vue d'un rendu sur casque.

III.2.1 Protocole expérimental : méthode " MUSHRA modifiée "

La méthode expérimentale d'évaluation utilise principalement le paradigme de la recommandation UIT-R BT.500 [IUT, 2012] MUSHRA et prend en compte la méthodologie proposée par Le Bagousse [Le Bagousse and Paquier M., 2011, Le Bagousse et al., 2012, Le Bagousse, 2014], détaillée en I.6.

Dans cette étude, nous souhaitons comparer quatre méthodes de spatialisation, afin de déterminer laquelle est la mieux adaptée pour la représentation binaurale. Trois axes perceptifs ont été évalués selon la méthode de Le Bagousse, *la qualité globale*, la représentation spatiale *espace* et *le timbre*. A la différence de la méthode d'évaluation de référence [Le Bagousse et al., 2012], nous considérons que les différents systèmes à évaluer ne dégradent pas le signal, ce qui a permis d'écarter de l'évaluation l'axe *défauts*.

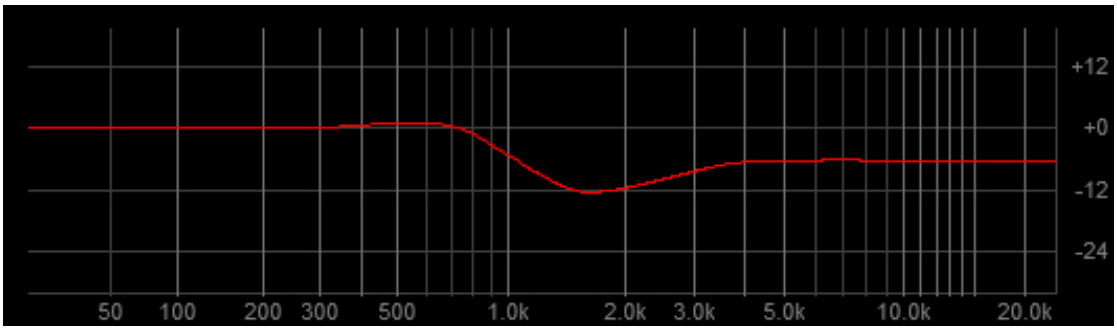


Figure III.3 : Réponse en fréquence (Module) du filtrage spectral utilisé pour créer l'ancre.

La notation se fait sur une échelle continue qui ne comporte aucun symbole ou label, afin d'encourager les auditeurs à utiliser l'echelle dans toute sa dynamique. Afin d'éviter toute inversion ou confusion, les labels *"meilleure qualité"* et *"moins bonne qualité"* sont affichés aux extrémités. Ces valeurs limites de l'échelle correspondent à 0 et 1 pour l'analyse numérique des résultats.

Aucun signal de référence n'a été introduit, afin de ne pas orienter les réponses des auditeurs. D'autre part, au lieu de trois ancrages séparés, tel que proposé par la méthode de Le Bagousse, nous avons souhaité l'introduction d'un ancrage unique de basse qualité comportant simultanément des dégradations selon les trois axes perceptifs que nous souhaitons évaluer. Il est à préciser que ce choix a été effectué de manière arbitraire, afin de réduire le nombre de stimuli et donc la durée du test. Des études ultérieures de Le Bagousse [Le Bagousse, 2014] ont démontré que le regroupement des ancrages dans un ancrage unique ne modifie pas la réponse des auditeurs.

Le but des ancrages est de fixer la plus basse qualité à atteindre afin de définir une fourchette de la dynamique d'évaluation. Les dégradations introduites par Le Bagousse, dans les ancrages proposés respectifs, nous ont paru excessivement fortes et auraient provoqué une compression des évaluations dans le haut de l'échelle, compte tenu de la nature du test.

Or, le but du test était d'évaluer la qualité fournie par différentes méthodes de spatialisation sonore et non d'évaluer les dégradations induites par une compression du signal.

Une ancre spécifique aux besoins de cette étude a donc été définie à partir des signaux stéréophoniques issus du couple AB.

Tout d'abord, pour introduire une dégradation du timbre, un filtrage spectral a été effectué. Il s'agit d'un filtre passe-bas, de fréquence de coupure 700 Hz, de pente -10 dB/oct, suivi d'un plateau à -10 dB à partir de 1.5 KHz (figure III.3). Ce filtre a été implémenté à l'aide d'un plugin VST.

Il a été montré [Kozamernik et al., 2007] qu'une réduction de la largeur de l'image sonore comme ancrage spatial induit une notation d'environ 56/100 sur échelle MUSHRA, en contraste avec l'utilisation d'un signal monophonique, noté 45/100. Afin de conserver une dynamique importante dans l'évaluation du critère *espace*, l'ancre a été soumise à une réduction de l'image de la façon suivante,

$$Sa_l = 0.75Ss_l + 0.25Ss_r \,, \tag{III.1a}$$

$$Sa_r = 0.75Ss_r + 0.25Ss_l \,, \tag{III.1b}$$

où les signaux d'ancrage sont notés Sa et ceux issus de la paire stéréo Ss. Les indices r et l correspondent respectivement aux signaux du canal droite et gauche.

Nous avons considéré que les dégradations ainsi introduites affectent également la qualité globale perçue.

17 sujets ont pris part au test (7 experts et 10 naïfs). Tous les sujets sont habitués au passage de test audio et les sujets experts sont habitués à l'écoute du son 3D dans le cadre de leurs activités professionnelles.

Le test est composé de deux sessions pour une durée totale d'environ 2 heures pendant laquelle les sujets étaient libres de faire des pauses à leur gré. Les deux sessions étaient séparées d'une pause obligatoire d'au moins 10 minutes. Les sujets ont été gratifiés pour leur participation au test.

Il a été demandé à chaque sujet d'ajuster le niveau sonore lors de l'écoute du premier échantillon, au niveau d'écoute qu'il juge confortable, et de ne pas le modifier jusqu'à la fin du test.

Dans la première session, les sujets évaluent simultanément la qualité en termes de *timbre* et *espace* pour chaque extrait (figure III.4). Les différents extraits sont joués en ordre aléatoire. Lors de la deuxième session, les mêmes extraits sont joués dans un ordre aléatoire différent et les sujets évaluent la qualité globale.

L'interface de test a été développée sous Matlab (figure III.4), suivant les caractéristiques décrites précédemment et la recommandation [IUT, 2012]. L'interface ne permet pas le passage à l'échantillon suivant tant que tous les échantillons n'ont pas été écoutés au moins une fois et qu'au moins un curseur n'a été déplacé.

L'expérience a été menée en parallèle dans deux locaux isolés acoustiquement et équipés du même matériel de restitution à savoir,

- un casque fermé HFI-580 Ultrasone,
- une interface audio Terratec Phase 26, utilisant le protocole ASIO avec une fréquence d'échantillonnage à 48 kHz et une résolution de 24 bits,
- un ordinateur de contrôle.

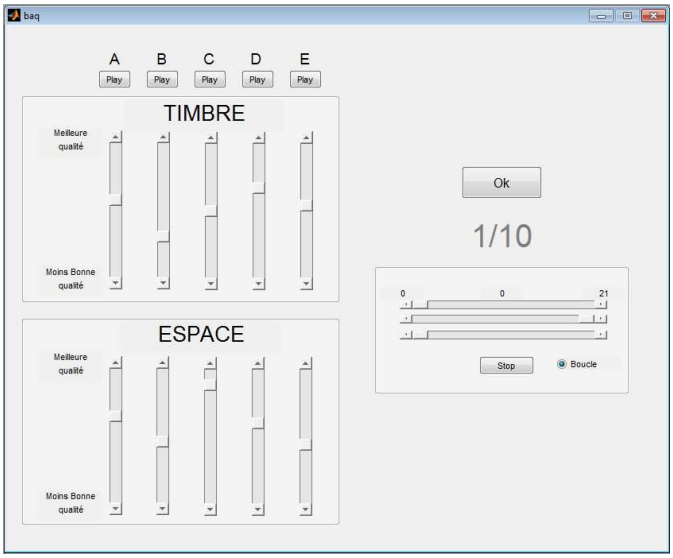


Figure III.4 : Interface de test.

III.2.2 Stimuli

Tableau III.1 : Liste des extraits audio.

Extrait No.	Titre	Type	Durée (s)
1	Ambiance	Public	13
2	Applaudissements	Public	14
3	Rires	Public + dialogue de comédiens	19
4	Musique	Orchestre	21
5	Dialogue	Dialogue de comédiens	18
6	Aria	Soprano Aria	16
7	Duo	Basse + Soprano	18
8	Trio	Ténor + Ténor + Basse	17
9	Quatuor	2 Ténors + 2 Soprani	20
10	Chœur	Chœur	24

Un ensemble d'échantillons a été extrait de l'enregistrement de la représentation publique de l'opéra de Mozart (tableau III.1). Ils constituent un échantillonnage représentatif des musiques et ambiances rencontrées pendant le spectacle.

Cinq versions sont comparées lors du test. Elles sont issues des dispositifs de captation décrits en III.1 et sont traitées comme détaillé en III.1.3, ainsi qu'un ancrage de faible qualité présenté dans III.2. Ces cinq versions sont identifiées par la suite avec les labels suivants :

- *"KU100"* : enregistrement binaural capté par la tête acoustique Neumann® KU100,
- *"Stereo"* : signal stéréophonique capté par la paire AB,
- *"SF dec 1"* : enregistrement ambisonique à l'ordre 1 comportant un décodage binaural classique (I.4.3.c),
- *"SF dec 2"* : enregistrement ambisonique à l'ordre 1 comportant le décodage binaural défini par Berge [Berge and Barrett, 2010] (I.5.2.b),
- *"Ancre"* : ancrage de basse qualité défini en III.2.

III.3 Résultats

Pour chaque version de spatialisation, 170 notes (10 extraits × 17 sujets) ont été obtenues, ceci pour chaque axe perceptif évalué (*qualité globale*, *timbre*, et *espace*). Pour l'ensemble des sujets et des échantillons, les ancres ont été reconnues avec une note moyenne de 0,005 (avec un intervalle de confiance à 95% de ± 0,002) pour la qualité globale et 0,01 (± 0,007) pour l'espace et le timbre. Ces résultats confirment que l'ancre choisie est bien appropriée pour les besoins de cette étude. Ils confirment également qu'elle peut être utilisée comme une mesure de la pertinence des réponses des sujets. La discrimination des sujets est effectuée comme suit : lorsqu'un sujet ne détecte pas correctement l'ancre de basse qualité, les résultats correspondant à l'échantillon en question sont supprimés, car cette erreur est sans doute liée à d'un manque d'attention temporaire ou à une erreur de manipulation. Si le même sujet ne parvient pas à détecter l'ancrage sur au moins deux échantillons, l'ensemble des résultats du sujet est écarté de l'analyse.

In fine, 12 échantillons et les résultats de quatre sujets, deux experts et deux naïfs ont été écartés. L'analyse est donc effectuée sur les 123 résultats des 13 sujets restants, pour l'évaluation de chaque axe perceptif.

Les figures III.5, III.6 et III.7 affichent les résultats de ce test. Les notes moyennes ainsi que les intervalles de confiance à 95% sont affichés (en ordonnée) pour les différentes "versions" (en abscisse). L'évaluation de chaque axe perceptif est présenté de façon indépendante. De plus, la figure III.5a présente la moyenne obtenue sur l'ensemble d'échantillons et des axes perceptifs. Les figures III.6 et III.7 affichent respectivement la moyenne pour chaque échantillon et chaque sujet sur chaque axe perceptif.

Sur la figure III.5a, on observe tout d'abord que les notes moyennes obtenues pour les différentes technologies de spatialisation se trouvent entre 0,4 et 0,7, ce qui confirme que l'ancre choisie ne compresse pas la dynamique des résultats. On observe également qu'en termes de "qualité globale", l'enregistrement binaural "KU100" est largement préféré, suivi du décodage binaural de l'ambisonique *"SF dec 2"*. La paire stéréophonique arrive en troisième position, ce qui prouve l'apport du binaural par rapport à la stéréo pour un rendu sur casque. En outre, tel qu'illustré en figure III.5, les sujets naïfs ont tendance à surévaluer la qualité et le timbre du décodage de l'ambisonique *"SF dec 2"* en la plaçant au même niveau que l'enregistrement binaural. Les experts sont plus critiques dans ces deux axes perceptifs et placent cette méthode au même niveau que l'enregistrement stéréo. Les deux catégories de sujets arrivent à un consensus lorsqu'il s'agit de la captation binaurale.

Les résultats obtenus sur l'ensemble des échantillons sont très corrélés entre eux avec

des "valeurs-P"[1] inférieures à 0,01 et des coefficients de corrélation très proches de 1 quel que soit l'axe perceptif. L'échantillon "2-Applaudissements" ne suit pas les mêmes tendances. En effet, il présente des valeurs-P toujours supérieures à 0,1 et des coefficients de corrélation inférieurs à 0,8. Les figures III.6 illustrent bien ces résultats, car l'ensemble des échantillons présentent la même allure globale à l'exception de l'échantillon n°2.

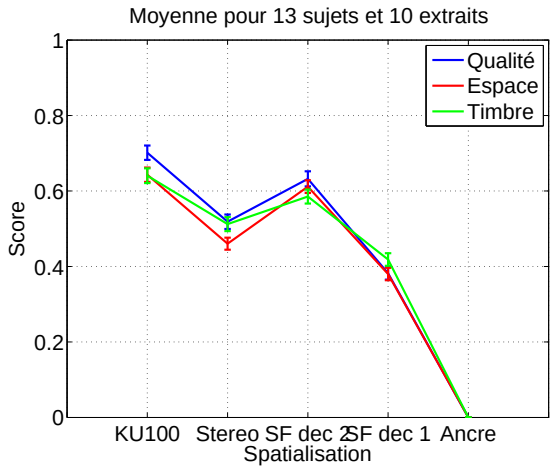
En effet, pour cet échantillon, la qualité de la prise stéréophonique est préférée à tous les autres systèmes. En termes d'espace, ce sont les deux décodages de l'ambisonique qui sont préférés. Enfin, en termes de timbre, le décodage classique de l'ambisonique *"SF dec 1"* et la stéréo se trouvent en tête et à des niveaux équivalents et la prise binaurale obtient un score inférieur à toutes les autres méthodes.

Les résultats obtenus pour cet échantillon peuvent s'expliquer par la compression dynamique du signal apporté au signal binaural durant cet extrait afin d'éviter la saturation.

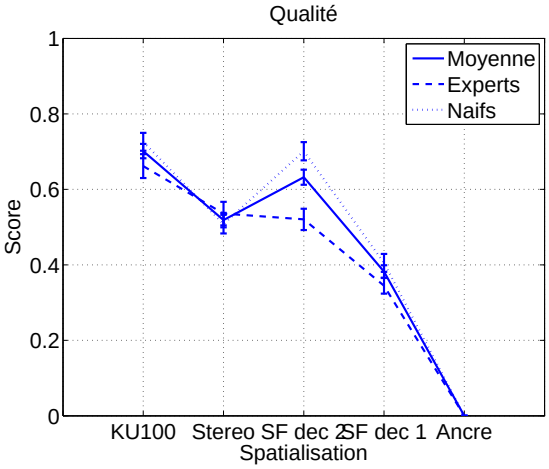
Sur la figure III.7 (réponses moyennes par sujet), l'allure générale des réponses est similaire, à l'exception de celles des sujets "5" et "10". En effet, la matrice de corrélation des résultats affiche des valeurs très proches de 1 pour l'ensemble des sujets, à l'exception de deux sujets ($r \simeq 0{,}7$ et $P > 0{,}1$).

Enfin, les résultats nous permettent de dire que les trois axes perceptifs sont corrélés entre eux. En effet, une analyse statistique permet d'obtenir des valeurs-p nulles et des coefficients de corrélation de 0,73 entre la qualité globale et qualité spatiale, et de 0,78 entre la qualité globale et la qualité du timbre. La valeur du coefficient de corrélation entre la qualité de timbre et l'espace est aussi de 0,78.

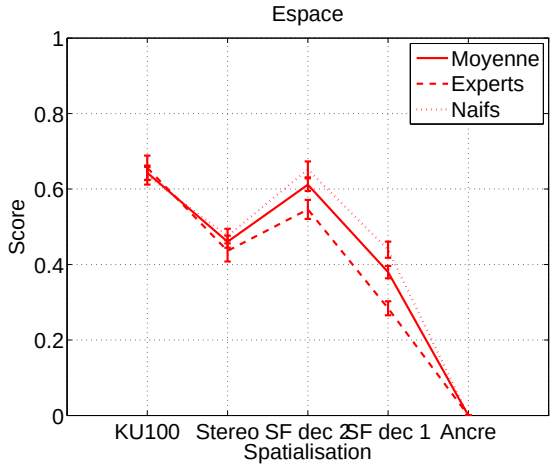
[1] Les "valeurs-P" permettent de considérér un résultat comme statistiquement significatif quand les valeurs sont très proches de 0 (<0,01) [Wasserman, 2004].



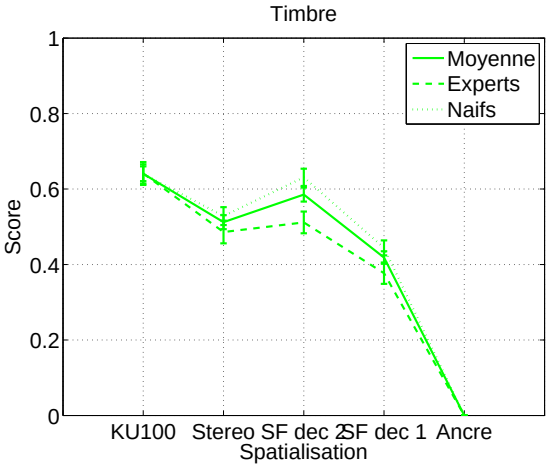
(a) Evaluation obtenue pour l'ensemble des sujets pour les trois axes perceptifs.



(b) Evaluation de la qualité par les 5 sujets experts et les 8 sujets naïfs, et moyenne de l'ensemble des sujets.

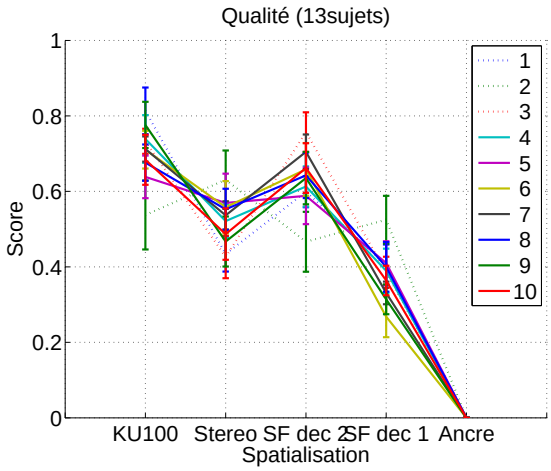


(c) Evaluation de l'espace par les 5 sujets experts et les 8 sujets naïfs, et moyenne de l'ensemble des sujets.

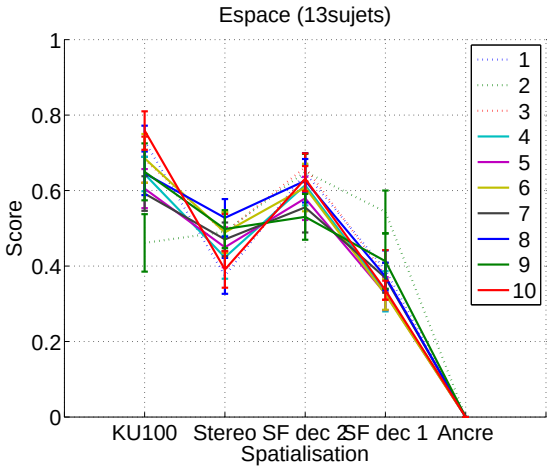


(d) Evaluation du timbre par les 5 sujets experts et les 8 sujets naïfs, et moyenne de l'ensemble des sujets.

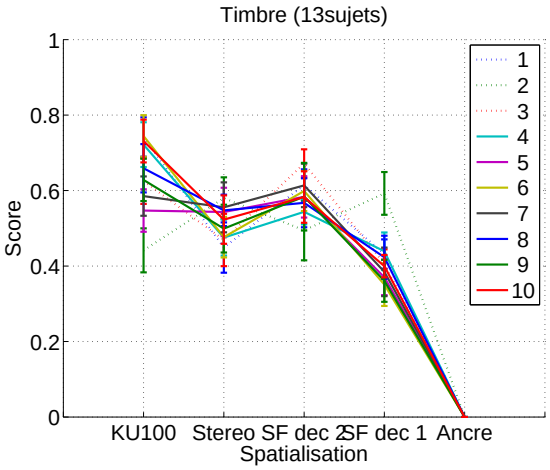
Figure III.5 : Moyenne et intervalle de confiance à 95% des notes obtenus pour les attributs perceptifs qualité, espace et timbre, en fonction de la méthode de captation et décodage. Calcul effectué sur 13 sujets et pour 10 extraits.



(a) Evaluation de l'attribut qualité.

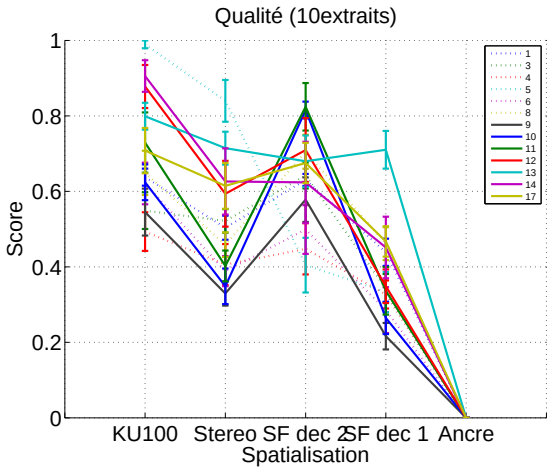


(b) Evaluation de l'attribut espace.

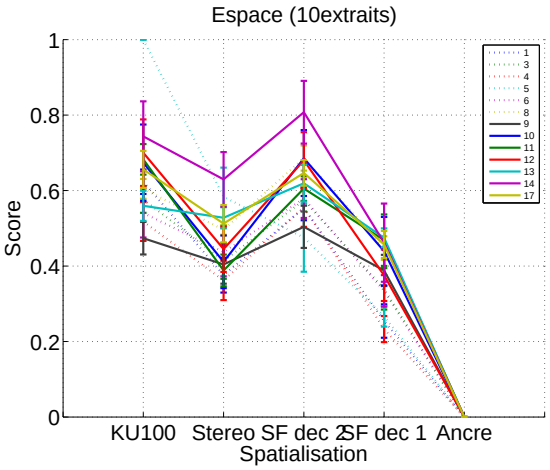


(c) Evaluation de l'attribut timbre.

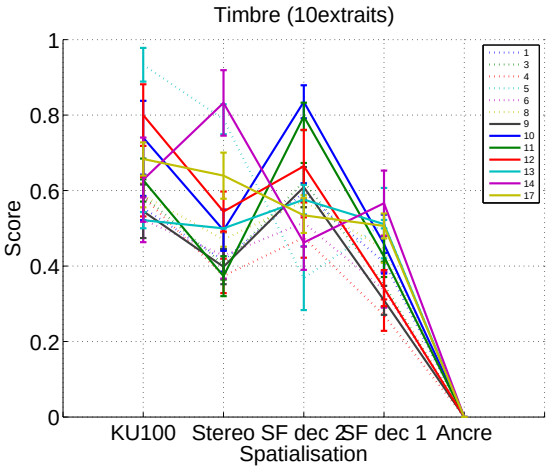
Figure III.6 : Moyenne et intervalle de confiance à 95% des notes obtenues pour chaque extrait, pour les attributs perceptifs qualité, espace et timbre, en fonction de la méthode de captation et décodage. Calcul effectué sur 13 sujets.



(a) Evaluation de l'attribut qualité.



(b) Evaluation de l'attribut espace.



(c) Evaluation de l'attribut timbre.

Figure III.7 : Moyenne et intervalle de confiance à 95% des notes obtenues pour chaque sujet, des attributs perceptifs, qualité, espace et timbre, en fonction de la méthode de captation et décodage. Calcul effectué sur 10 extraits.

III.4 Conclusions du test subjectif

L'expérience de la captation de l'opéra de Mozart, *L'Enlèvement au Sérail* à l'Opéra de Rennes, nous a permis de constituer une base de données, qui a été utilisée pour la création d'un corpus d'échantillons sonores représentatifs de ce genre d'événements, à partir de quatre techniques de captation sonore utilisées pour la restitution au casque.

Les résultats du test d'écoute ont montré que l'enrichissement spatial d'une scène sonore apporte autant à la qualité sonore que la qualité de restitution du timbre. L'étude permet de constater que la méthode de la captation binaurale est la plus adaptée pour la captation d'une scène sonore spatialisée avec un rendu sur casque. En particulier, les colorations apportées par la prise sonore binaurale ne sont pas perçues ici comme une dégradation de la qualité sonore. Au contraire, la prise de son binaurale permet un renforcement de la qualité spatiale et par conséquent de la qualité globale perçue.

Le test d'écoute a également mis en évidence que l'utilisation d'un microphone ambisonique couplé à un décodage binaural adapté permet d'atteindre des résultats comparables à ceux de la prise de son binaurale, offrant ainsi la possibilité d'une captation plus compatible avec les dispositifs mobiles, et permettant de s'affranchir des contraintes imposées par la captation binaurale directe.

Cette expérience suggère une piste possible pour la prise de son 3D avec un dispositif nomade, en s'inspirant du décodage actif de l'ambisonique. En effet, la qualité perçue de cette méthode est supérieure à celle issue du décodage classique de l'ambisonique, prouvant ainsi que la qualité du rendu spatialisé dépend autant de la mise en œuvre du décodage que du dispositif de captation utilisé.

Cette méthode est basée sur la localisation des sources sonores et est associée à un décodage "objet", tel que nous l'avons évoqué précédemment. Ce type de méthode offre la possibilité de pouvoir utiliser différents jeux d'HRTF et permet le rendu sur différents types de dispositifs de restitution.

Afin d'explorer la piste ainsi évoquée, nous proposons, dans les chapitres IV et V, une méthode permettant de réduire le nombre de capteurs mis en jeu pour ce type de captation.

Avant d'écarter la piste du décodage ambisonique conventionnel, il est nécessaire de comprendre les raisons des mauvais résultats perceptifs obtenus avec cette technique. Une évaluation objective des performances de ce décodage est proposée en III.5, en s'intéressant principalement à la projection des HRTF sur les harmoniques sphériques.

III.5 Évaluation objective

Les résultats obtenus lors de l'évaluation perceptive presentée en III.3 ont mis en évidence deux éléments clefs pour la suite. Tout d'abord, l'utilisation d'un décodage adapté de l'ambisonique permet d'atteindre des résultats très proches des résultats obtenus avec une prise de son binaurale, et d'autre part, il a été démontré que le décodage classique de l'ambisonique ne donne pas les résultats escomptés en termes de qualité perçue. Afin d'approfondir nos conclusions sur ce type de décodage et de comprendre les paramètres affectant le signal, un protocole expérimental a été mis en place.

III.5.1 Protocole expérimental

Les HRTF utilisées pour la synthèse binaurale sont soumises à différénts pré-traitements comme il est décrit dans I.4.2.e. Nous allons donc nous intéresser au rôle qu'ils jouent dans le décodage ambisonique.

Nous utilisons ici la base de données des HRTF de Pernaux [Pernaux, 2003]. Le choix s'est porté sur cette base, car elle comporte un échantillonnage spatial régulier sur la partie supérieure de la sphère (élévations supérieures à -56,25°) et comporte 965 directions mesurées. La fréquence d'échantillonnage des réponses impulsionnelles est de 48 kHz sur 512 échantillons temporels pour les huit sujets constituant cette base. L'ensemble des analyses détaillées par la suite est arbitrairement issu du sujet n°1.

Pour évaluer l'impact des différents pré-traitements, la décomposition en harmoniques sphériques a été effectuée sur des HRTF ne comportant aucun pré-traitement (000) et comportant une combinaison de pré-traitements, conformément au tableau III.2.

Pré-traitement	Pré-traitement		
	Phase minimale + ITD	Lissage Fréq.	Interpolation
000			
100	*		
020		*	
120	*	*	
123	*	*	*

Tableau III.2 : Liste des pré-traitements appliqués.

L'encodage et le décodage ambisonique sont appliqués sur l'ensemble des directions des HRTF (965 directions mesurées et 1026 directions après interpolation[2]) , suivant les procédures décrites en I.4.3.b, I.4.3.c et I.5.2.a.

La projection sur les harmoniques sphériques (équations (I.16) et (I.17)) a été tronquée aux ordres 1, 4 et 30. Ce choix arbitraire a été effectué afin de se limiter aux ordres des dispositifs disponibles dans le commerce : le 1er ordre correspondant au microphone Soundfield® [Gerzon, 1975] et l'ordre 4 correspondant à l'Eigenmike®. Enfin, l'ordre 30 a été retenu afin de représenter une restitution complète, tel que décrit en I.4.3.c.

Cette évaluation se déroule en 2 parties :

- l'étude du comportement du décodage ambisonique à l'ordre 1 pour 4 haut-parleurs virtuels disposés selon des géométries différentes ;
  - **régulière sur la sphère :** sur les sommets d'un tétraèdre,
  - **régulière sur un cercle :** uniformement distribués sur le cercle décrit par un plan découpant la sphère,
- l'étude du comportement du décodage ambisonique aux ordres 1, 4 et 30 pour les 956 directions de HRTF disponibles.

III.5.1.a Critères d'analyse

La projection des HRTF sur les harmoniques sphériques à un ordre M implique une approximation de la représentation de cette fonction. Dans les études précédentes, l'erreur d'approximation des troncatures ambisoniques a principalement été analysée à partir de l'énergie des fonctions projetées [Moreau, 2006] [Bamford, 1995], [Daniel, 2001] [Poletti, 2000].

L'analyse effectuée ici porte principalement sur la restitution de l'information spatiale contenue dans les HRTF. En effet, l'analyse énergétique d'approximation des HRTF délivre uniquement une note moyenne de la qualité de reconstruction sur l'ensemble des directions (C.4). Afin de compléter cette information, il est nécessaire :

- de définir des critères permettant de fournir une information précise sur les indices de localisation (binauraux et monoauraux définis en I.2) contenus dans les HRTF,
- de mesurer comment ces indices sont conservés ou dégradés par l'approximation apportée par la projection sur des harmoniques sphériques.

[2] Afin de compléter le maillage sur la partie manquante de la sphère.

L'indicateur ILD (I.2.1) est calculé pour chaque direction (θ, ϕ) par la méthode proposée par Larcher [Larcher, 2001] suivant

$$ILD = 10 \log_{10} \frac{\int_{1.5kHz}^{10kHz} |HRTF_l(f)|^2 \, \mathrm{d}f}{\int_{1.5kHz}^{10kHz} |HRTF_r(f)|^2 \, \mathrm{d}f} \,, \tag{III.2}$$

où l et r désignent les oreilles gauche et droite respectivement.

L'indicateur ITD est quant à lui calculé selon la méthode de Nam [Nam et al., 2008] (I.4.2.e). L'analyse du rendu spectrale est effectuée à l'aide du critère *Inter-Subject Spectral Difference* (ISSD) introduit par Middlebrooks [Middlebrooks, 1999] [3] . Cette méthode est basée sur le calcul de la variance de la différence entre le spectre original $HRTF(f)$ et le spectre reconstruit $\widehat{HRTF}(f)$ pour chaque direction (θ, ϕ) entre 4 kHz et 13 kHz, selon la relation,

$$ISSD = \frac{1}{9kHz} \int_{4kHz}^{13kHz} \left(10 \log_{10} \frac{\widehat{HRTF}(f)}{HRTF(f)} - \Psi \right)^2 \mathrm{d}f \,, \tag{III.3}$$

où

$$\Psi = \frac{1}{9kHz} \int_{4kHz}^{13kHz} 10 \log_{10} \frac{\widehat{HRTF}(f)}{HRTF(f)} \, \mathrm{d}f \,. \tag{III.4}$$

Afin d'obtenir un indice unique pour un ensemble de HRTF, la valeur moyenne l'ISSD est calculée sur l'ensemble des directions. Selon Middelbrooks [Middlebrooks, 1999], une valeur d'ISSD inférieure à 6,18 dB² représente une restitution optimale de l'ensemble des HRTF considérées d'un point de vue perceptif.

[3] Il est a remarquer que la valeur de l's'exprime en dB² car il s'agit d'une variance.

III.5.2 Résultats expérimentaux

III.5.2.a Décodage ambisonique à l'ordre 1 sur 4 haut-parleurs virtuels

Tout d'abord, nous avons limité les directions des HRTF à 4 afin d'effectuer une restitution parfaite de l'ambisonique à l'ordre 1 (I.4.3.b).

Géométrie régulière sur la sphère : dans un premier temps, les directions ont été choisies sur les sommets d'un tétraèdre, afin d'obtenir une configuration idéale lors de l'inversion de la matrice de décodage (I.4.3.c), grâce à l'équation (I.25).

La position du haut-parleur n est donnée en coordonnées sphériques par le vecteur H_n, suivant

$$H_1 = \begin{pmatrix} 0 \\ -\frac{\pi}{6} \\ 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} \frac{2\pi}{3} \\ -\frac{\pi}{6} \\ 1 \end{pmatrix}, \quad H_3 = \begin{pmatrix} -\frac{2\pi}{3} \\ -\frac{\pi}{6} \\ 1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} 0 \\ \frac{\pi}{2} \\ 1 \end{pmatrix}. \tag{III.5}$$

Tel qu'exprimé dans les tableaux III.3, III.4 et III.5 ainsi que sur les figures III.8 et III.9, l'ensemble d'indicateurs, ITD, ILD, ISSD est parfaitement reconstruit dans cette configuration. Les erreurs affichées sont introduites par les pré-traitements effectués et ne dépendent pas du décodage ambisonique.

Pré-traitement	θ (rad)	0	$2\pi/3$	$-2\pi/3$	0	$\overline{ISSD}$	σ
	ϕ	$-\pi/6$	$-\pi/6$	$-\pi/6$	$\pi/2$		
000	PT	0	0	0	0	0	0,0
	amb.	0	0	0	0	0	0,0
100	PT	0	0	0	0	0	0,0
	amb.	0	0	0	0	0	0,0
020	PT	1	0	0	0	0	0,3
	amb.	1	0	0	0	0	0,3
120	PT	1	0	0	0	0	0,3
	amb.	1	0	0	0	0	0,3

Tableau III.3 : Évaluation de la reconstruction spectrale (ISSD en dB^2) obtenue par décodage ambisonique à l'ordre 1 en utilisant 4 haut-parleurs virtuels disposés sur les sommets d'un tétraèdre : comparaison du pré-traitement seul (PT) et du décodage ambisonique après pré-traitement (amb.).

Pré-traitement	θ (rad)	0	$2\pi/3$	$-2\pi/3$	0	Δ ITD (ms)	
	ϕ	$-\pi/6$	$-\pi/6$	$-\pi/6$	$\pi/2$	$\bar{\Delta}$	σ
ITD cible (Original)		0,02	-0,54	0,52	0,02	-	-
000	PT	0,02	-0,54	0,52	0,02	0,00	0,00
	amb.	0,02	-0,54	0,52	0,02	0,00	0,00
100	PT	0,00	-0,51	0,50	0,00	0,02	0,00
	amb.	0,00	-0,51	0,50	0,00	0,02	0,00
020	PT	0,02	-0,54	0,52	0,02	0,00	0,00
	amb.	0,02	-0,54	0,52	0,02	0,00	0,00
120	PT	0,00	-0,51	0,50	0,00	0,02	0,00
	amb.	0,00	-0,51	0,50	0,00	0,02	0,00

Tableau III.4 : ITD (en ms) obtenue par décodage ambisonique à l'ordre 1 en utilisant 4 haut-parleurs virtuels disposés sur les sommets d'un tétraèdre : comparaison du pré-traitement seul (PT) et du décodage ambisonique après pré-traitement (amb.).

Pré-traitement		rad θ	$2\pi/3$	$-2\pi/3$	0	Δ ILD (dB)	
		ϕ	$-\pi/6$	$-\pi/6$	$-\pi/6$	$\pi/2$	$\bar{\Delta}$	σ
ILD cible (Original)			-0,3	14,5	-12,3	1,1	-	-
000	PT		-0,3	14,5	-12,3	1,1	0,0	0,0
	amb.		-0,3	14,5	-12,3	1,1	0,0	0,0
100	PT		-0,3	14,5	-12,3	1,1	0,0	0,0
	amb.		-0,3	14,5	-12,3	1,1	0,0	0,0
020	PT		-0,2	14,3	-12,3	1,1	0,1	0,1
	amb.		-0,2	14,3	-12,3	1,1	0,1	0,1
120	PT		-0,2	14,3	-12,3	1,1	0,1	0,1
	amb.		-0,2	14,3	-12,3	1,1	0,1	0,1

Tableau III.5 : ILD (en dB) obtenue par décodage ambisonique à l'ordre 1 en utilisant 4 haut-parleurs virtuels disposés sur les sommets d'un tétraèdre : comparaison du pré-traitement seul (PT) et du décodage ambisonique après pré-traitement (amb.).

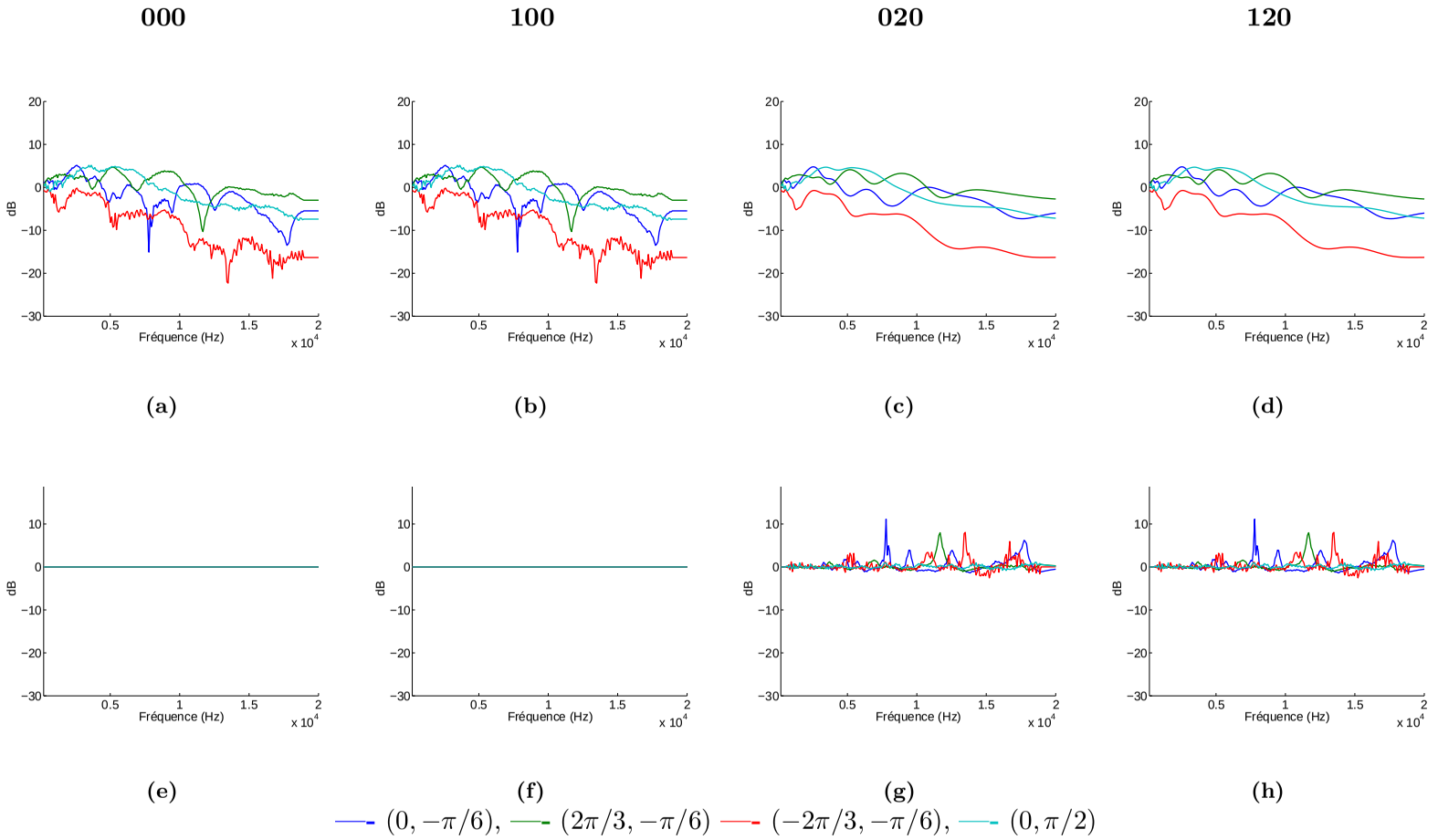


Figure III.8 : Spectre d'amplitude des 4 HRTF (ligne 1 (a)(b)(c)(d)) et erreur associée (ligne 2 (e)(f)(g)(h)) pour les 4 pré-traitements 000, 100, 020 et 120. Les directions des HRTF décrivent un tétraèdre.

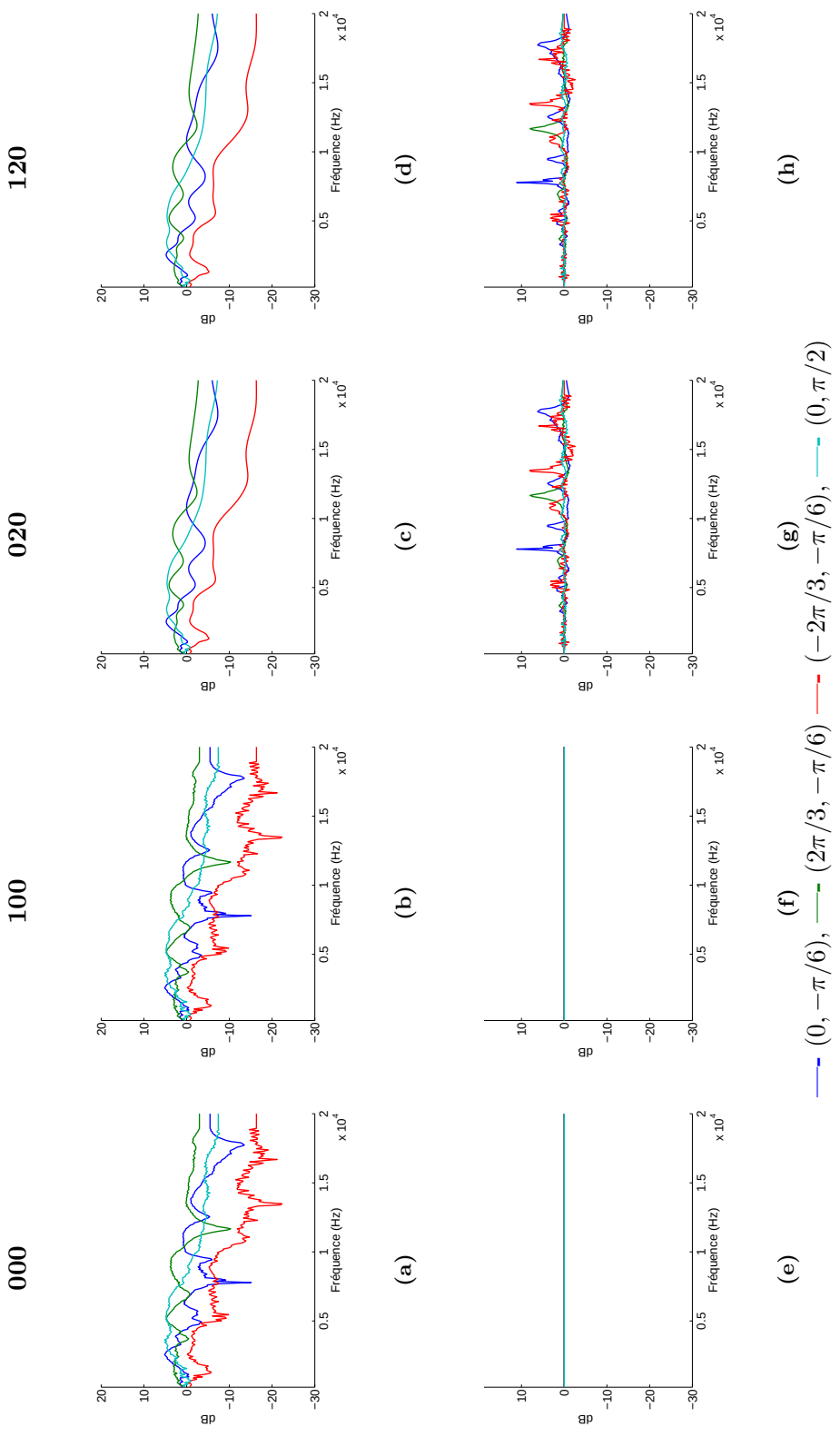


Figure III.9 : Spectre d'amplitude de la reconstruction ambisonique des 4 HRTF (ligne 1 (a)(b)(c)(d)) et erreur associée (ligne 2 (e)(f)(g)(h)) pour les 4 pré-traitements 000, 100, 020 et 120. Les directions des HRTF décrivent un tétraèdre.

Géométrie irrégulière sur la sphère : Un deuxième décodage est effectué sur un ensemble de 4 haut-parleurs disposés de façon irrégulière sur la sphère. Afin d'illustrer la dépendance du décodage sur la position des haut-parleurs virtuels, ils sont placés de façon arbitraire sur un plan coupant la sphère et distribués de façon équidistante sur le cercle résultant.

La position des haut-parleurs n est donnée en coordonnées sphériques par les vecteurs H_n, suivant

$$H_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} \pi \\ 0 \\ 1 \end{pmatrix}, \quad H_3 = \begin{pmatrix} \frac{\pi}{2} \\ -\frac{\pi}{4} \\ 1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} \frac{3\pi}{2} \\ \frac{\pi}{4} \\ 1 \end{pmatrix}. \tag{III.6}$$

La première ligne de la figure III.10 affiche le spectre d'amplitude des HRTF des directions retenues après les pré-traitements. La deuxième ligne de la même figure affiche l'écart d'amplitude qui en résulte. En comparant cette erreur avec l'erreur introduite par le décodage ambisonique sur la figure III.11, on constate que dans cette configuration, le décodage ambisonique introduit des erreurs importantes. L'indicateur ISSD affiche des valeurs très élevées (tableau III.6). Selon le critère fixé par Middelbrooks [Middlebrooks, 1999] (qui précise une erreur <6,18 dB^2), des HRTF présentant des ISSD aussi élevées ne sont pas exploitables pour la restitution binaurale.

L'indicateur ILD est également dégradé (tableau III.7). Il est surestimé pour les HRTF disposées sur le plan médian et sous-estimé pour les autres directions.

D'autre part, le décodage ambisonique dans cette configuration ne permet pas la reconstruction de l'ITD comme on vérifie dans le tableau III.8 .

Ces résultats laissent entrevoir l'attention particulière à apporter à la position des haut-parleurs virtuels dans le décodage binaural de l'ambisonique. En effet, l'équation (I.4.3.c) effectue une approximation lors de l'inversion de la matrice, qui est très sensible à ces paramètres.

Pré-traitement		rad θ					
	θ	0	π	$\pi/2$	$3\pi/2$		
	ϕ	0	0	$-\pi/4$	$\pi/4$	$\overline{ISSD}$	σ
000	PT	0	0	0	0	0	0,0
	amb.	1	13	9	26	12	10,2
100	PT	0	0	0	0	0	0,0
	amb.	1	30	14	17	15	12,1
020	PT	0	1	0	0	0	0,3
	amb.	2	11	8	22	11	8,7
120	PT	0	1	0	0	0	0,3
	amb.	1	34	13	18	16	13,5

Tableau III.6 : Évaluation de la reconstruction spectrale (ISSD en dB^2) obtenue par décodage ambisonique à l'ordre 1 en utilisant 4 haut-parleurs virtuels placés aux directions indiquées dans l'équation III.6 : comparaison du pré-traitement seul (PT) et du décodage ambisonique après pré-traitement (amb.).

Pré-traitement		rad θ				Δ ILD (dB)	
	θ	0	π	$\pi/2$	$3\pi/2$		
	ϕ	0	0	$-\pi/4$	$\pi/4$	$\bar{\Delta}$	σ
ILD cible (Original)		-0,4	0,2	17,0	-11,4	-	-
000	PT	-0,4	0,2	17,0	-11,4	0,0	0,0
	amb.	-1,8	0,6	3,8	-10,2	4,0	6,1
100	PT	-0,4	0,2	17,0	-11,4	0,0	0,0
	amb.	-2,2	-1,8	8,5	-10,9	3,2	3,6
020	PT	-0,4	0,3	17,0	-11,5	0,0	0,0
	amb.	-1,7	0,5	3,5	-9,9	4,1	6,2
120	PT	-0,4	0,3	17,0	-11,5	0,0	0,0
	amb.	-2,2	-1,8	8,5	-11,0	3,2	3,6

Tableau III.7 : ILD (en dB) obtenue par décodage ambisonique à l'ordre 1 (amb.) en utilisant 4 haut-parleurs virtuels placés aux directions indiquées dans l'équation III.6 : comparaison du pré-traitement seul (PT) et du décodage ambisonique après pré-traitement (amb.).

Pré-traitement		rad				Δ ITD (ms)	
	θ	0	π	$\pi/2$	$3\pi/2$		
	ϕ	0	0	$-\pi/4$	$\pi/4$	$\bar{\Delta}$	σ
ITD cible (Original)		0,02	-0,01	-0,39	0,47	-	-
000	PT	0,02	-0,01	-0,39	0,47	0,00	0,00
	amb.	0,02	0,12	-0,09	0,01	0,23	0,20
100	PT	0,00	0,00	-0,38	0,46	0,01	0,00
	amb.	0,00	0,00	-0,06	0,00	0,21	0,23
020	PT	0,01	-0,01	-0,40	0,48	0,00	0,00
	amb.	0,02	0,13	-0,09	0,00	0,23	0,20
120	PT	0,00	0,00	-0,38	0,46	0,01	0,00
	amb.	0,00	0,00	-0,06	0,00	0,21	0,23

Tableau III.8 : ITD (en *ms*) obtenue par décodage ambisonique à l'ordre 1 en utilisant 4 haut-parleurs virtuels placés aux direction indiquées dans l'équation III.6 : comparaison du pré-traitement seul (PT) et du décodage ambisonique après pré-traitement (amb.).

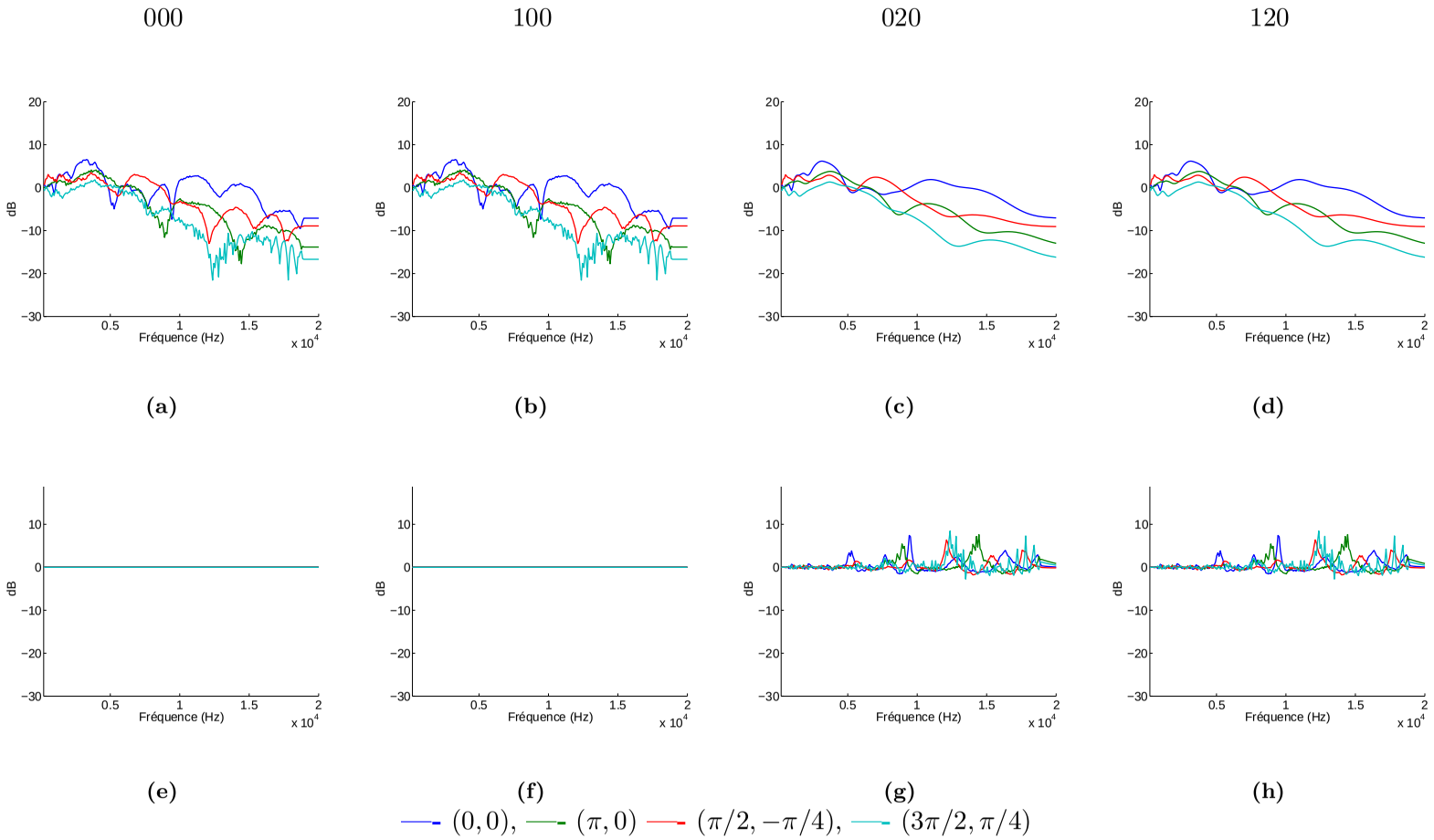


Figure III.10 : Spectre d'amplitude des 4 HRTF (ligne 1 (a)(b)(c)(d)) et erreur associée (ligne 2 (e)(f)(g)(h)) pour les 4 pré-traitements 000, 100, 020 et 120. Les directions des HRTF sont uniformément distribuées sur un cercle.

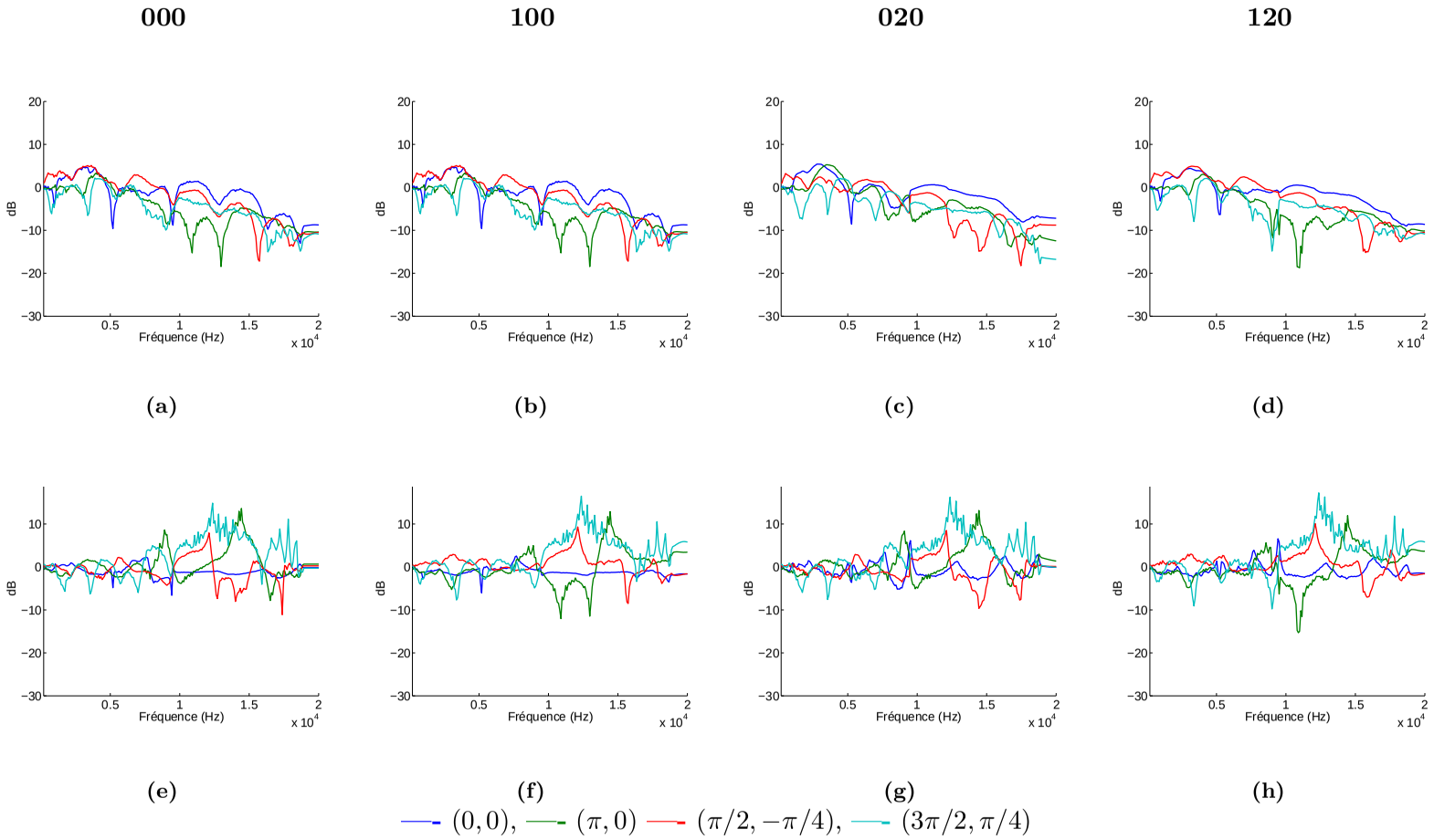


Figure III.11 : Spectre d'amplitude de la reconstruction ambisonique des 4 HRTF (ligne 1 (a)(b)(c)(d)) et erreur associée (ligne 2 (e)(f)(g)(h)) pour les 4 pré-traitements 000, 100, 020 et 120. Les directions des HRTF sont uniformément distribuées sur un cercle.

III.5.2.b Décodage ambisonique aux ordres 1, 4 et 30 de l'ensemble de HRTF

D'une part, la figure III.12 affiche le spectre d'amplitude des HRTF comportant uniquement les pré-traitements, ainsi que leur écart par rapport aux HRTF non traitées. D'autre part, les figures III.13, III.14 et III.15 affichent le spectre d'amplitude des HRTF après reconstruction ambisonique et après pré-traitement, ainsi que les erreurs associées. Les HRTF affichées correspondent à celles de l'oreille gauche. On observe que les erreurs sont principalement situées du côté droit de la tête (contralatéral), car la dynamique des HRTF est réduite de ce côté. Les poids attribués lors de la projection sur des harmoniques sphériques sont de moindre importance pour ces valeurs, dégradant ainsi leur représentation.

A l'ordre 1 (figure III.13), le décodage ambisonique effectue un lissage spatial très important, en supprimant toute information contenue dans les hautes fréquences. A l'ordre 4 (figure III.14), le lissage est moins important. La distribution spatiale des harmoniques sphériques introduit des vallées qui peuvent modifier les indices spectraux et dégrader notamment la perception de l'élévation. En effet l'énergie est concentrée à proximité des lobes principaux des harmoniques sphériques utilisés. A l'ordre 30 (figure III.15), la reconstruction du spectre d'amplitude est bien effectuée, y compris aux hautes fréquences, et les erreurs obtenues se confondent avec celles introduites par les pré-traitements effectués (figure III.12).

Reconstruction spectrale (ISSD) Les observations effectuées sur le spectre d'amplitude sont confortées par l'ISSD obtenue. La figure III.16 affiche l'ISSD obtenue pour chacune des directions des HRTF suite aux pré-traitements et au décodage ambisonique. L'ISSD est très dégradée, notamment pour les directions contralatérales et pour les ordres de reconstruction 1 et 4, les valeurs dépassent largement la limite définie par Middelbrooks [Middlebrooks, 1999] (i.e 6,18 dB^2). A l'ordre 30, seules quelques erreurs persistent, notamment pour les HRTF ne comportant aucun pré-traitement (figure III.16(l)).

Les valeurs moyennes des ISSD obtenues ont été reportées dans le tableau III.9. L'utilisation des pré-traitements ne dégradent pas la représentation spatiale (ISSD < 1 dB^2) (figure III.16(a,b,c)) ce qui justifie leur utilisation. Conformément aux résultats affichés sur les figures III.16(d) à (k), les valeurs de l'ISSD dépassent le seuil de 6,18 dB^2. En revanche, la reconstruction ambisonique à l'ordre 30 donne des valeurs d'ISSD inférieures à cette valeur. Cependant, l'écart-type de ISSD de la reconstruction ambisonique des HRTF non traitées affiche une valeur de 10 dB^2. Cette valeur contraste avec celle des HRTF comportant un traitement (σ ISSD < 2.5 dB^2) En d'autres termes, l'utilisation d'une HRTF simplifiée (sans ITD) augmente donc considérablement la qualité de la restitution ambisonique. De la même manière, les HRTF lissées sont mieux reconstruites

que les HRTF non lissées.

Dans le cas présent, l'apport de l'interpolation sur des directions non mesurées est négligeable pour la reconstruction ambisonique. Ce traitement est assez coûteux du point de vue des calculs, et dans le cas présent, peut être écarté. Néanmoins, son intérêt reste à prouver. En effet, le jeu d'HRTF utilisé est assez complet et uniformément distribué sur les directions existantes. Il est possible que ce pré-traitement prenne tout son sens dans l'utilisation d'une base de données présentant un échantillonnage plus parcimonieux.

Ordre	pré-traitement	ISSD en dB²		
		$\bar{X}$	σ	dégradation moyenne
traitement seul	[0 0 0]	0,0	0,0	0,00
	[1 0 0]	0,0	0,0	0,00
	[1 2 0]	0,5	0,6	0,00
	[1 2 3]	0,5	0,6	0,00
1	[0 0 0]	19,1	19,3	19,05
	[1 0 0]	21,3	22,8	21,28
	[1 2 0]	21,7	23,1	21,19
	[1 2 3]	21,4	22,4	20,96
4	[0 0 0]	18,6	18,3	18,57
	[1 0 0]	13,4	16,2	13,44
	[1 2 0]	13,9	17,0	13,43
	[1 2 3]	13,8	17,0	13,32
30	[0 0 0]	3,6	10,1	3,64
	[1 0 0]	0,5	1,8	0,53
	[1 2 0]	1,0	2,0	0,54
	[1 2 3]	1,2	2,3	0,70

Tableau III.9 : Évaluation de la reconstruction spectrale (ISSD) pour les différents ordres de l'ambisonique et différents pré-traitements. $\bar{X}$ représente la moyenne sur l'ensemble des directions et σ l'écart-type. La dégradation moyenne est calculée comme la différence entre $\bar{X}$ obtenue avant et après reconstruction ambisonique.

Reconstruction des indices binauraux (ITD, ILD) Tel que présenté dans le tableau III.10, l'ITD est relativement bien reconstruit à l'ordre 30. L'erreur suit une distribution gaussienne, de valeur moyenne proche de 15 μs et un écart-type de 3 μs. Pour les ordres inférieurs (1 et 4) et pour tous les types de pré-traitements, l'ITD est complètement ignoré avec une valeur moyenne oscillant autour de 0 s. Les figures III.20 affichent des erreurs du même ordre de grandeur que la valeur l'ITD pour les ordres 1 et 4.

Considérant que l'information sur le niveau est codée par le spectre (figure III.15), l'indicateur ILD est bien reconstruit pour l'ordre 30 comme affiché dans le tableau III.11 et conformément à ce que nous venons d'observer pour l'ISSD. L'erreur obtenue est

Ordre	pré-traitement	Δ ITD en ms		
		$\bar{X}$	σ	dégradation moyenne
traitement seul	[0 0 0]	0,00	0,00	0,00
	[1 0 0]	0,02	0,00	0,00
	[1 2 0]	0,02	0,00	0,00
	[1 2 3]	0,02	0,00	0,00
1	[0 0 0]	0,31	0,19	0,31
	[1 0 0]	0,33	0,20	0,31
	[1 2 0]	0,33	0,20	0,31
	[1 2 3]	0,33	0,20	0,31
4	[0 0 0]	0,24	0,19	0,24
	[1 0 0]	0,28	0,18	0,26
	[1 2 0]	0,29	0,18	0,27
	[1 2 3]	0,29	0,18	0,28
30	[0 0 0]	0,00	0,00	0,00
	[1 0 0]	0,02	0,00	0,00
	[1 2 0]	0,02	0,00	0,00
	[1 2 3]	0,02	0,00	0,00

Tableau III.10 : Δ ITD entre l'ITD obtenu pour des différents ordres de l'ambisonique et différents pré-traitements et l'ITD extrait des HRTF d'origine. $\bar{X}$ représente la moyenne sur l'ensemble des directions et σ l'écart-type. La dégradation moyenne est calculée comme la différence entre $\bar{X}$ des HRTF obtenues avant et après reconstruction ambisonique.

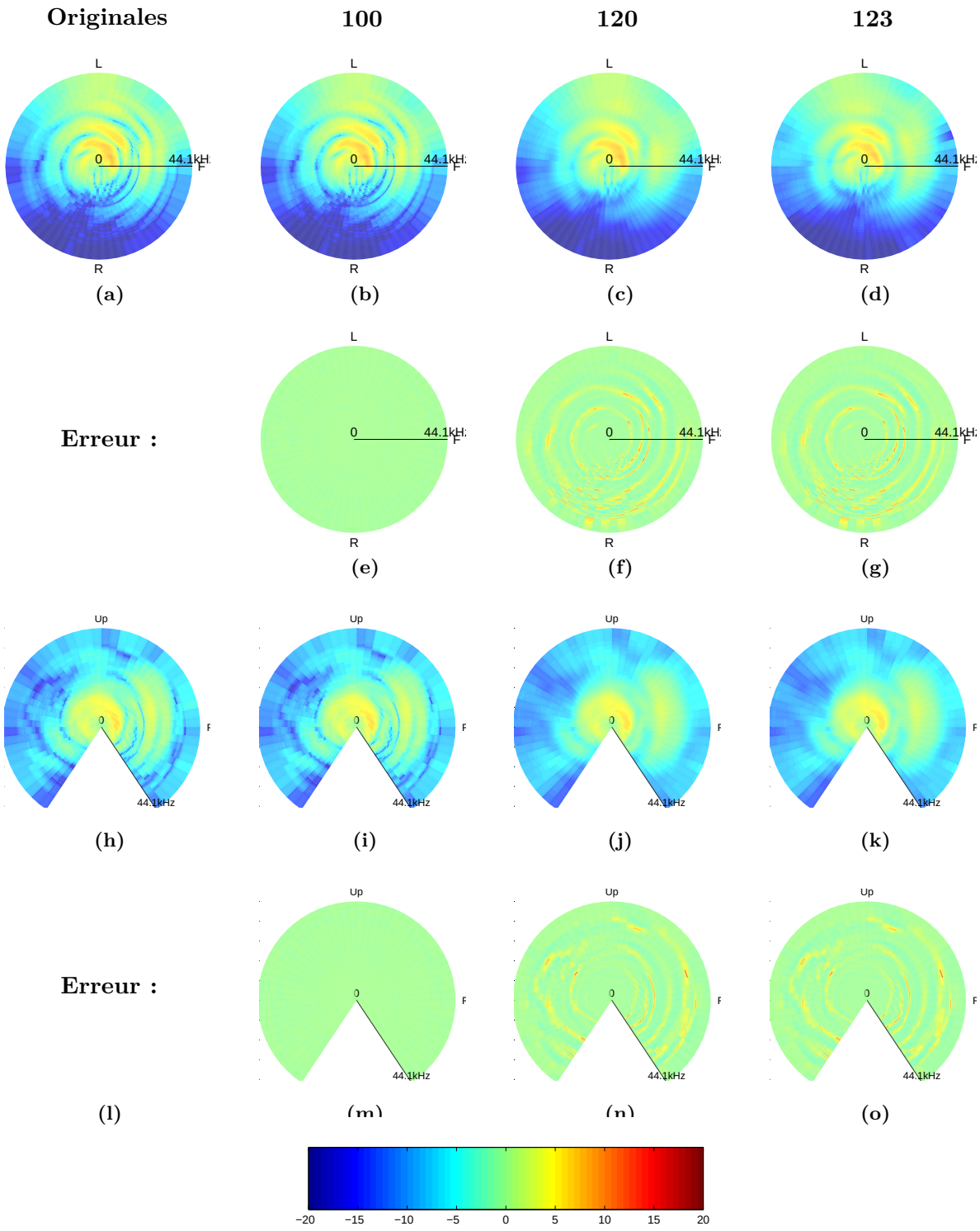
toujours inférieure à 1 dB.

Pour les ordres 1 et 4, la reconstruction de ILD pour des HRTF sans aucun pré-traitement présente une distribution gaussienne, de valeur moyenne voisine de 0 dB, avec un écart type de 3 dB. Pour quelques directions, la valeur maximale de l'erreur atteint 8 dB. Dans tous les cas, la variation spatiale est cohérente avec l'ILD naturel (figure III.18). Les résultats obtenus laissent apparaître que l'utilisation de tous les pré-traitements augmente l'erreur sur l'ILD à l'hémisphère nord, où l'ILD est surestimé.

Généralement, l'ILD des HRTF sans pré-traitements est bien reconstruit à l'ordre 1. On remarque quelques surestimations sur des secteurs situés autour de $(\frac{\pi}{2},-\frac{\pi}{8})$ et $(-\frac{\pi}{2},-\frac{\pi}{8})$. L'utilisation de pré-traitements augmente le nombre de zones de surestimation de l'ILD. Ceci s'explique notamment par la concentration de l'énergie autour des vecteurs principaux des harmoniques sphériques du premier ordre.

Ordre	pré-traitement	Δ ILD en dB		
		$\bar{X}$	σ	dégradation moyenne
traitement seul	[0 0 0]	0,00	0,00	0,00
	[1 0 0]	0,00	0,00	0,00
	[1 2 0]	0,06	0,05	0,00
	[1 2 3]	0,06	0,05	0,00
1	[0 0 0]	4,27	3,02	4,27
	[1 0 0]	3,30	2,86	3,30
	[1 2 0]	3,36	2,91	3,31
	[1 2 3]	3,47	2,94	3,41
4	[0 0 0]	2,43	1,89	2,43
	[1 0 0]	7,39	3,84	7,39
	[1 2 0]	7,59	3,89	7,53
	[1 2 3]	7,46	3,98	7,40
30	[0 0 0]	0,09	0,12	0,09
	[1 0 0]	0,07	0,08	0,07
	[1 2 0]	0,09	0,09	0,03
	[1 2 3]	0,12	0,11	0,06

Tableau III.11 : Δ ILD entre l'ILD obtenu pour des différents ordres de l'ambisonique et différents pré-traitements et ILD des HRTF d'origine. $\bar{X}$ représente la moyenne sur l'ensemble des directions et σ l'écart-type. La dégradation moyenne est calculée comme la différence entre $\bar{X}$ des HRTF obtenues avant et après reconstruction ambisonique.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o)

Figure III.12 : Spectre d'amplitude des HRTF sur le plan horizontal (ligne 1) et sur le plan médian (ligne 3) comportant les différents pré-traitements (colonnes) et les erreurs qui en résulte (lignes 2 et 4).

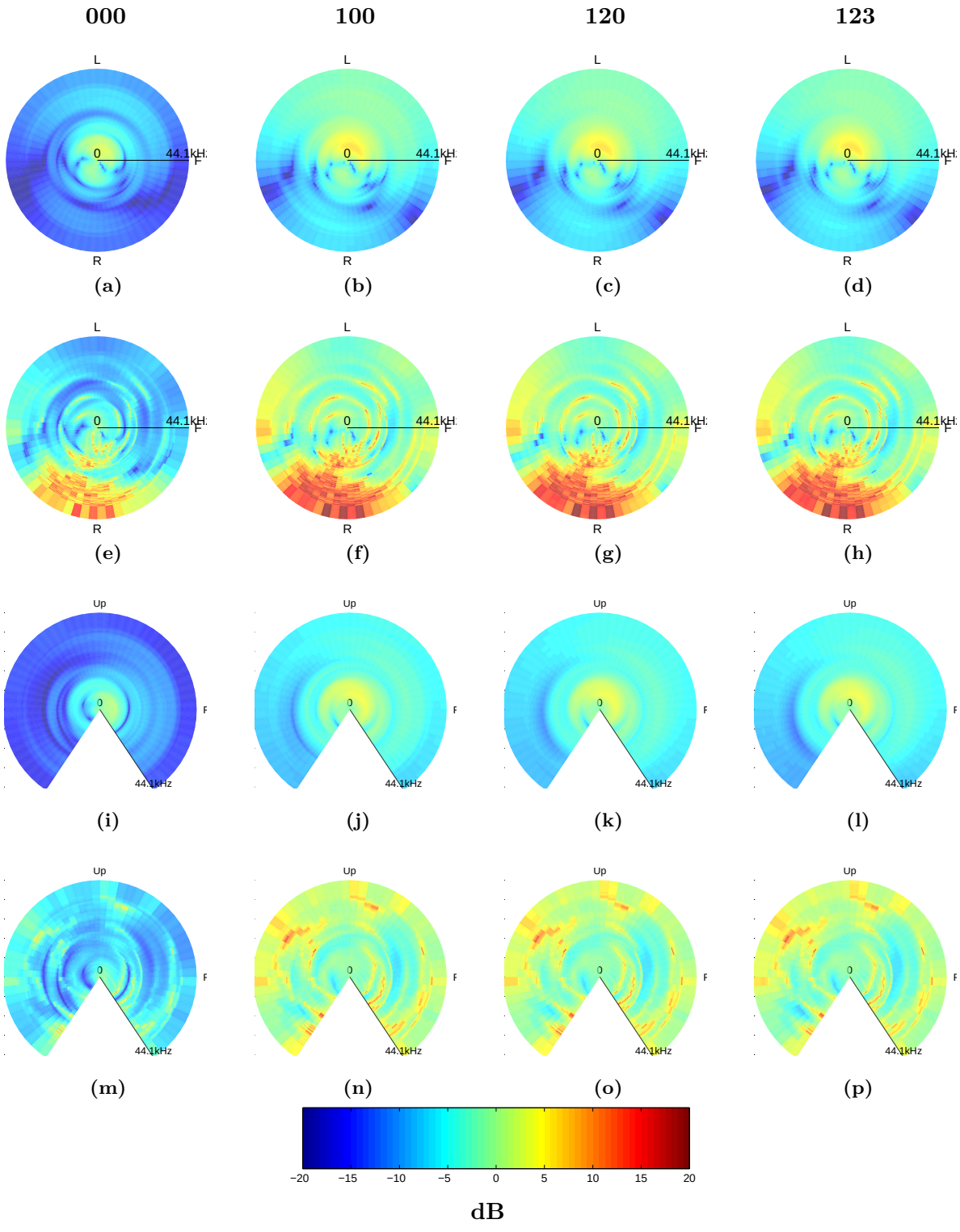


Figure III.13 : Spectre d'amplitude des HRTF sur le plan horizontal (ligne 1) et sur le plan médian (ligne 3) comportant les différents pré-traitements (colonnes) après **décodage ambisonique à l'ordre 1**. Erreurs associées en prenant comme référence les HRTF originales non traitées (lignes 2 et 4).

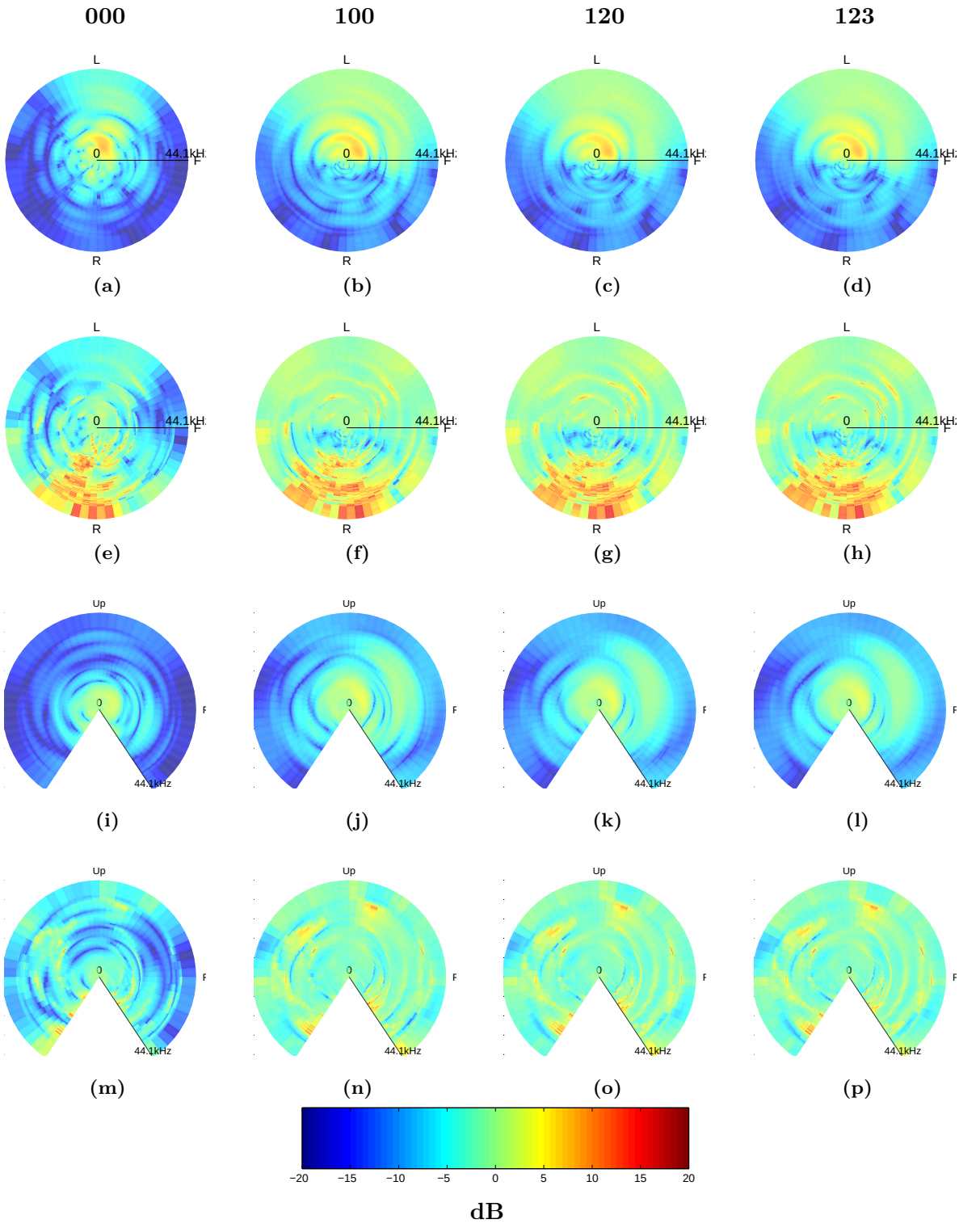


Figure III.14 : Spectre d'amplitude des HRTF sur le plan horizontal (ligne 1) et sur le plan médian (ligne 3) comportant les différents pré-traitements (colonnes) après **décodage ambisonique à l'ordre 4**. Erreurs associées en prenant comme référence les HRTF originales non traitées (lignes 2 et 4).

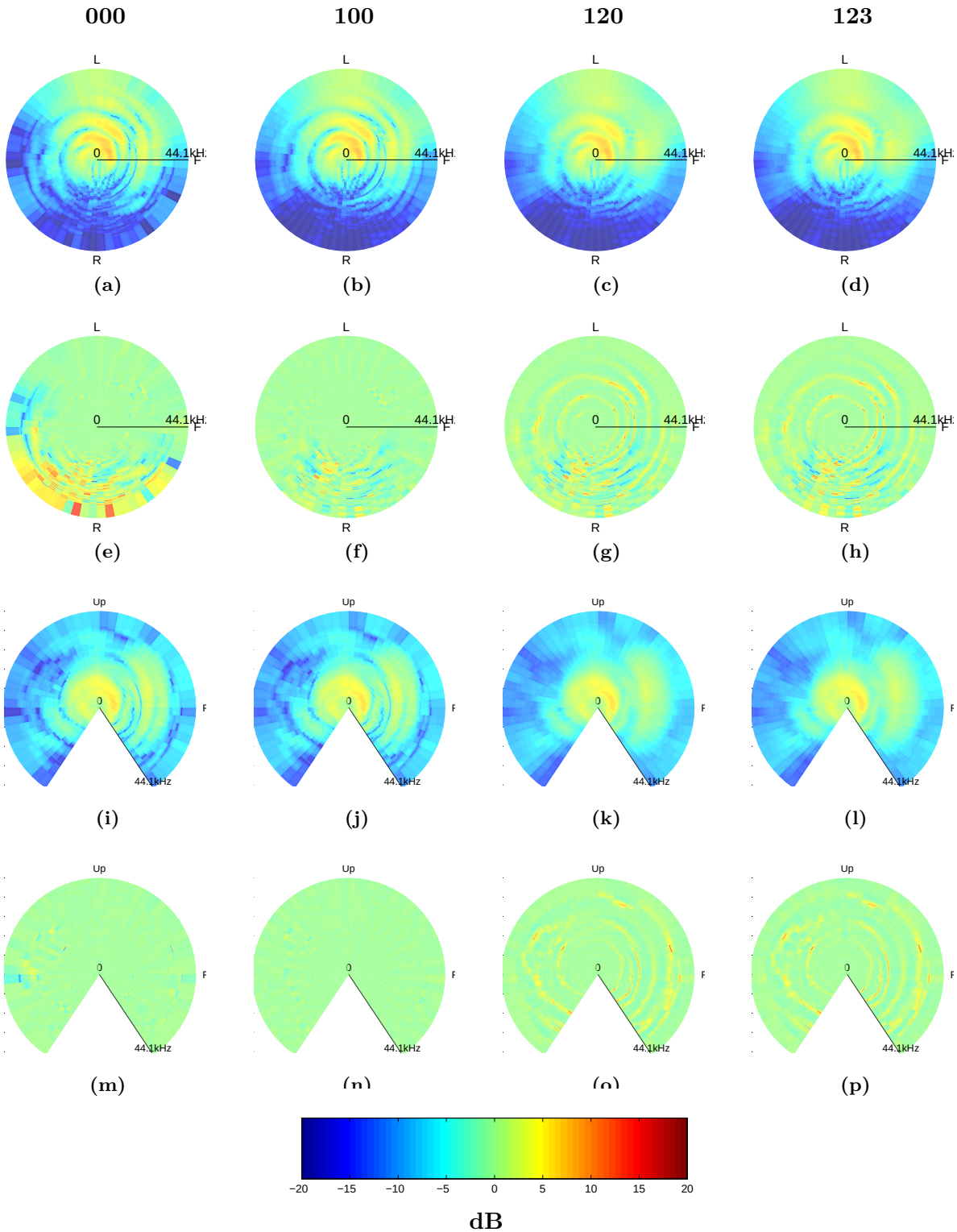


Figure III.15 : Spectre d'amplitude des HRTF sur le plan horizontal (ligne 1) et sur le plan médian (ligne 3) comportant les différents pré-traitements (colonnes) après **décodage ambisonique à l'ordre 30**. Erreurs associées en prenant comme référence les HRTF originales non traitées (lignes 2 et 4).

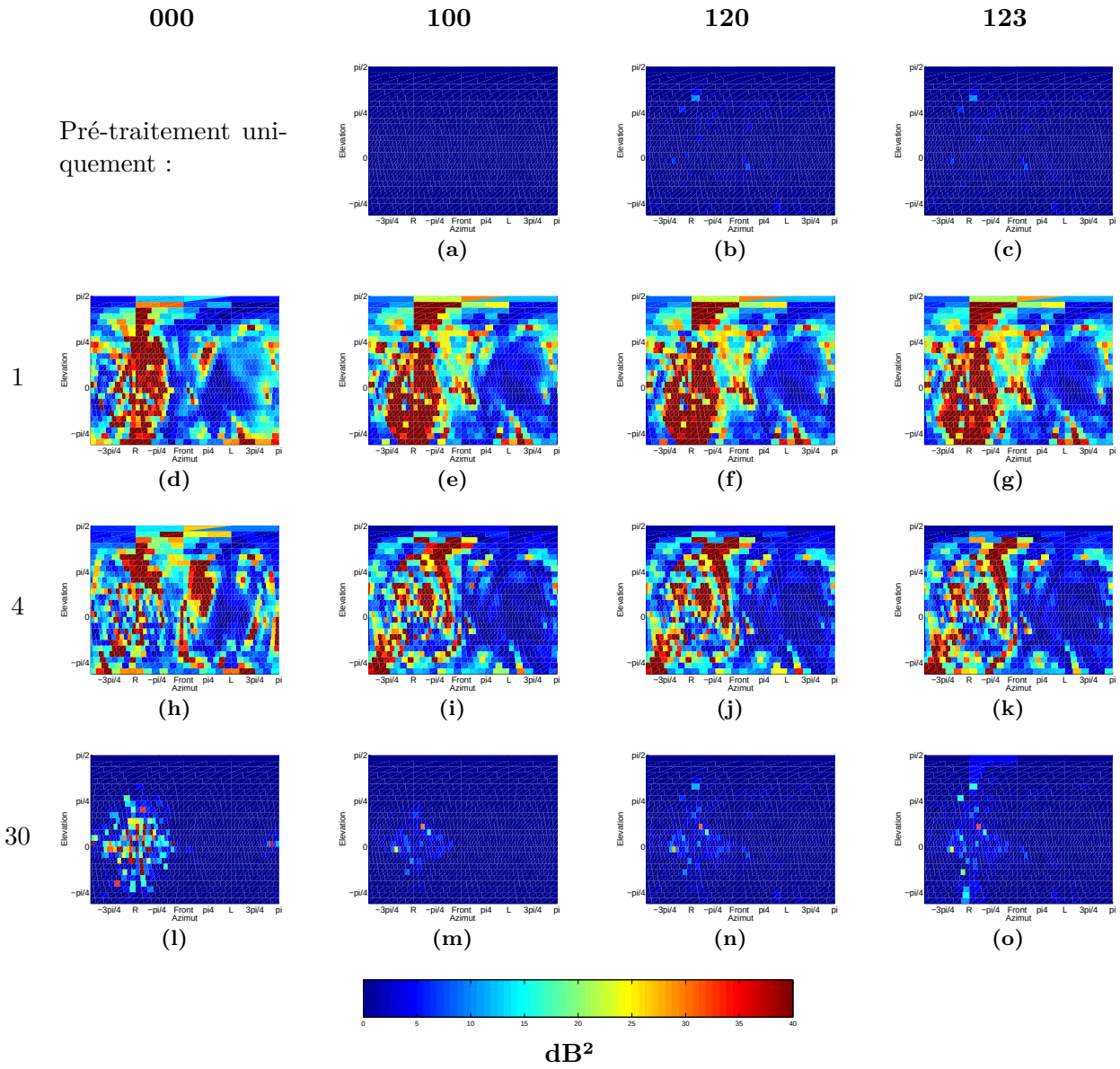


Figure III.16 : ISSD évaluée après le décodage ambisonique à l'ordre 1 (ligne 1), 4 (ligne 3) et 30 (ligne 5) de l'ensemble d'HRTF, en fonction des pré-traitements (colonnes). ISSD évaluée en prenant comme référence les HRTF originales non traitées

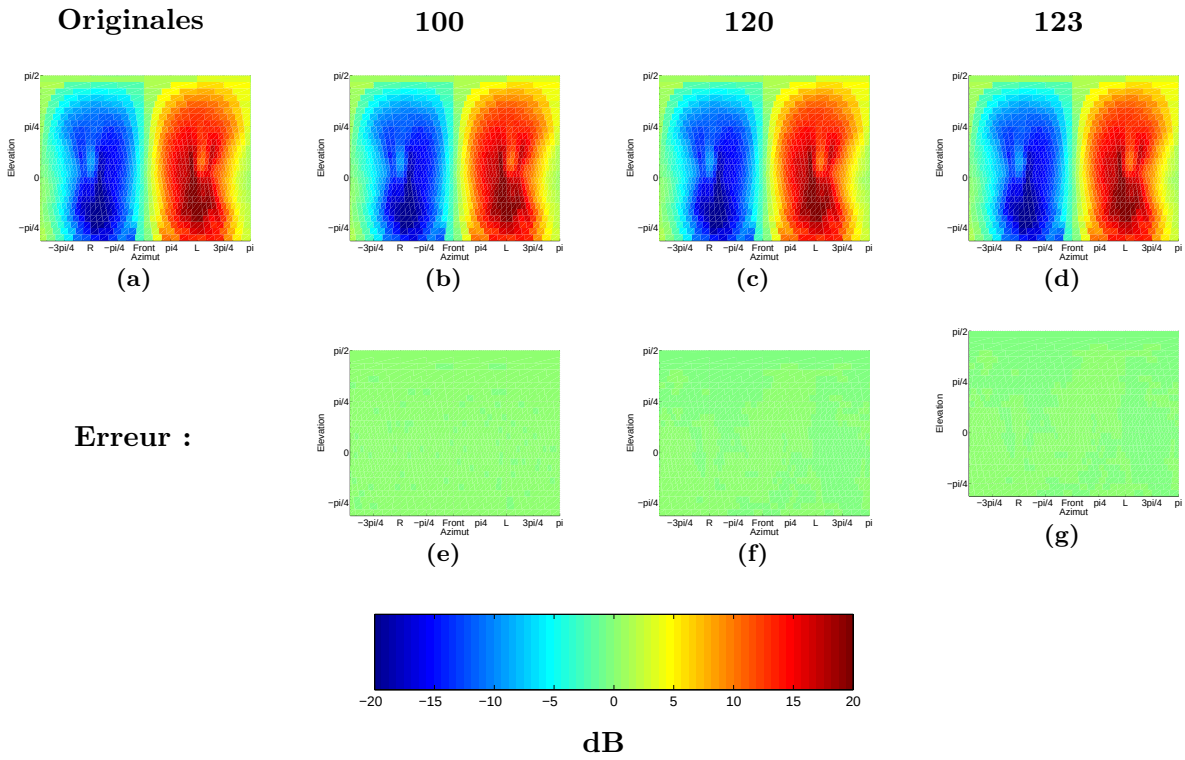


Figure III.17 : ILD évalué sur l'ensemble des directions des HRTF (ligne 1) comportant les différents pré-traitements (colonnes). Erreurs associées en prenant comme référence ILD évaluée sur l'HRTF originale (ligne 2).

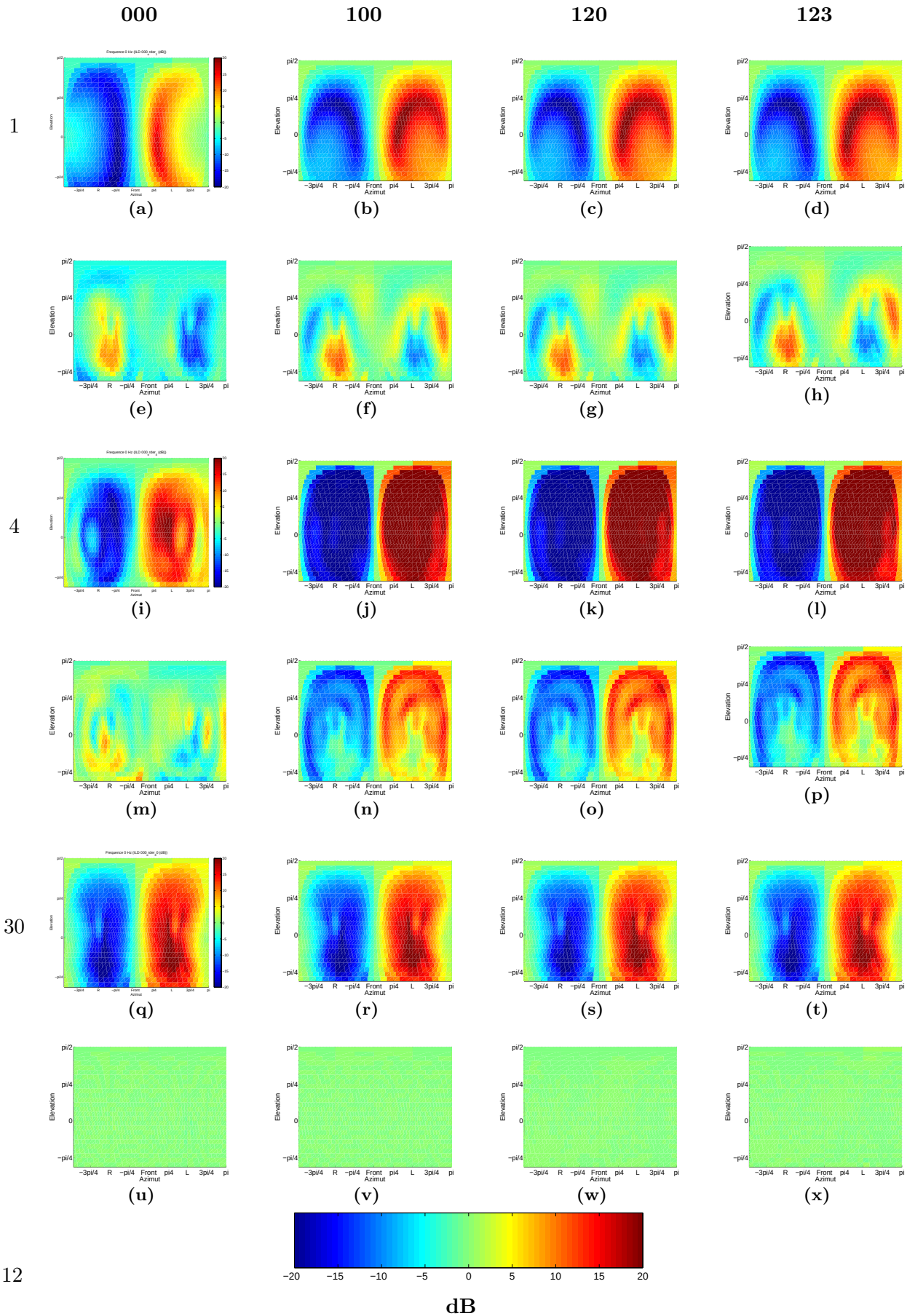


Figure III.18 : ILD résultant après le décodage ambisonique à l'ordre 1 (ligne 1), 4 (ligne 3) et 30 (ligne 5) de l'ensemble d'HRTF comportant les différents pré-traitements (colonnes). Erreurs associées en prenant comme référence ILD évaluée sur l'HRTF originale (lignes 2,4,6).

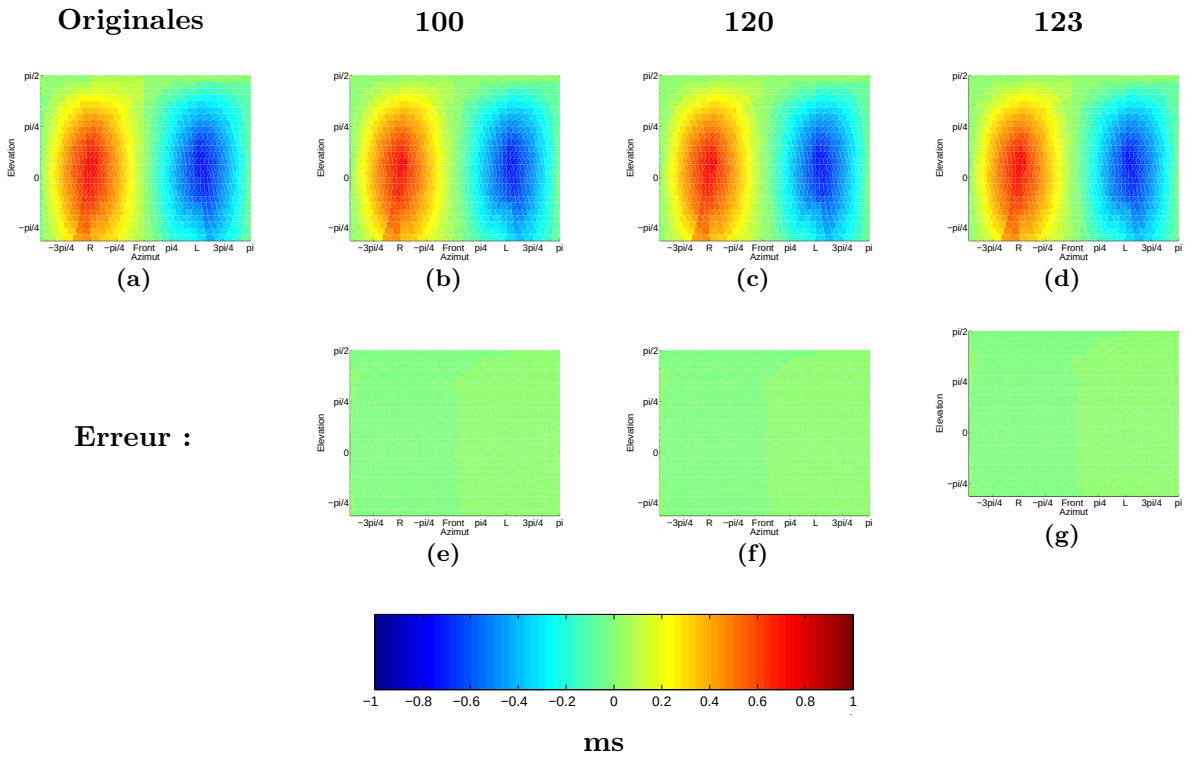


Figure III.19 : ITD évaluée sur l'ensemble des directions d'HRTF (ligne 1) comportant les différents pré-traitements (colonnes). Erreurs associées en prenant comme référence ITD évaluée sur l'HRTF originale (ligne 2).

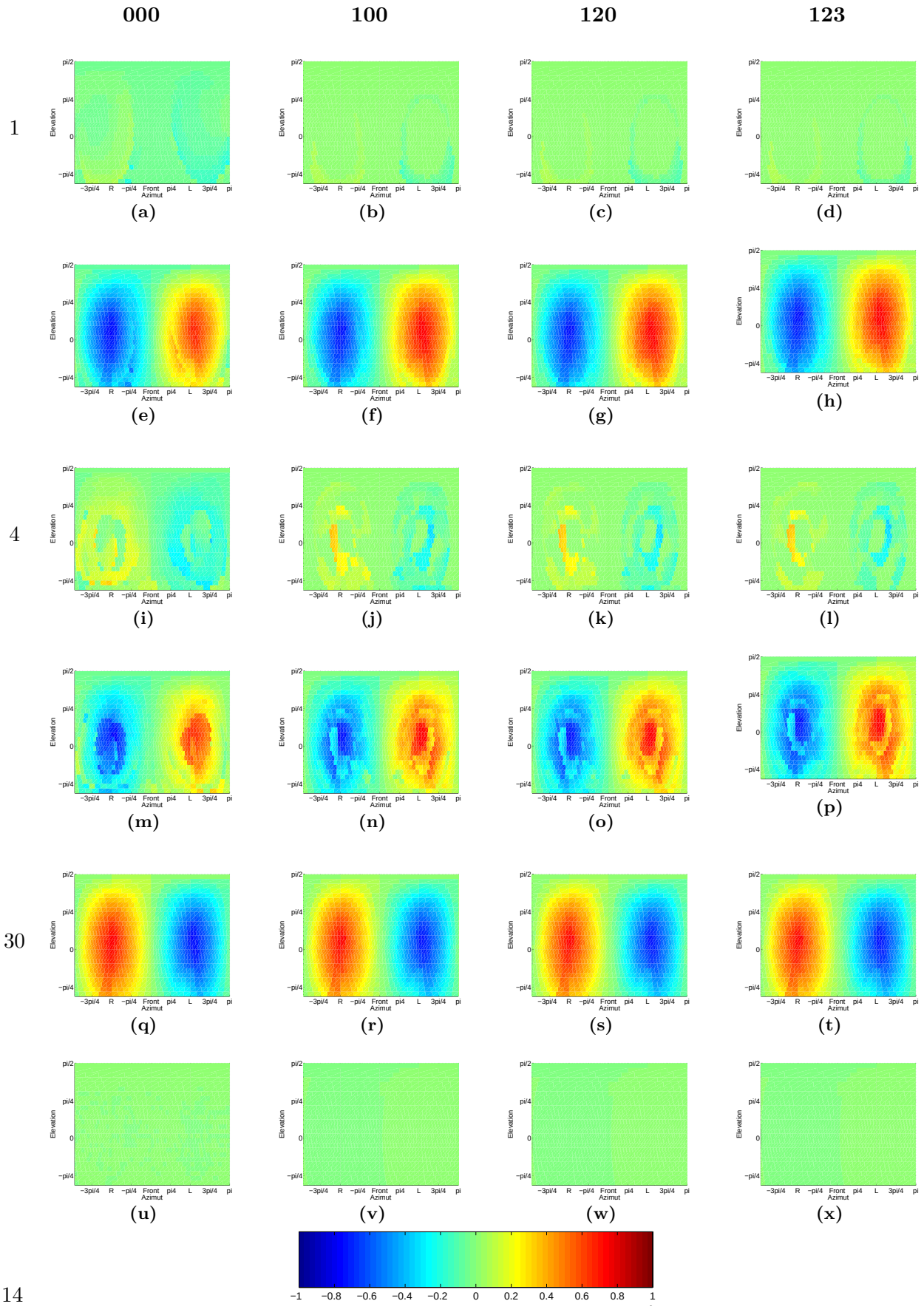


Figure III.20 : ITD évaluée après le décodage ambisonique à l'ordre 1 (ligne 1), 4 (ligne 3) et 30 (ligne 5) de l'ensemble d'HRTF comportant les différents pré-traitements (colonnes). Erreurs associées en prenant comme référence ITD évaluée sur l'HRTF originale (lignes 2,4,6).

III.6 Conclusion

L'étude du décodage binaural de l'ambisonique permet de mettre en évidence qu'il est très sensible aux différents paramètres utilisés pour sa mise en œuvre. Tout d'abord, l'inversion matricielle demande une distribution homogène des sources virtuelles sur la sphère, faute de quoi les indices permettant la perception spatiale sont fortement dégradés. Dans cette optique, la base des HRTF doit être suffisamment complète pour pouvoir effectuer le décodage des HRTF placées aux positions correctes.

Les résultats obtenus montrent également qu'une simplification des HRTF d'origine apporte une meilleure qualité de décodage, via notamment l'utilisation de filtres à phase minimale accompagnés d'un retard pur.

Dans l'approche classique, la projection des HRTF sur des harmoniques sphériques dégrade leur restitution aux premiers ordres, permettant d'écarter son utilisation dans une prise de son ambisonique. Dans cette approche, l'utilisation d'une base d'HRTF détaillée perd tout son sens. Les efforts de mise en œuvre pour l'obtenir ne sont pas perdus par le lissage spatial.

Les résultats de l'évaluation objective ne font que conforter le choix effectué dans la première partie de ce chapitre. Pour mémoire, le test d'écoute met en évidence que l'utilisation des décodages actifs donne des résultats comparables à la prise de son binaurale. Il est donc possible, en utilisant des approches hybrides, de s'affranchir des limitations imposées par le décodage classique en utilisant une phase d'analyse de la scène sonore dans le but de restituer au mieux l'ensemble des informations spatiales contenues dans les HRTF. C'est cette approche que nous avons retenue dans la suite de notre étude.

IV Proposition d'un prototype de prise de son 3D pour terminal mobile

IV.1 Stratégie

Dans l'approche de captation de l'image sonore 3D, une première phase de localisation des sources est nécessaire. Dans cette première phase, nous cherchons à extraire un signal représentatif de la scène sonore, afin de pouvoir effectuer un codage directionnel de celle-ci.

Nous allons donc étudier comment le choix des microphones, ainsi que leur configuration géométrique, peuvent apporter ces informations, tout en conservant l'objectif d'un système composé de moins de 4 capsules en association avec un algorithme adapté de localisation. Ce parti pris est adopté afin de proposer une solution comportant moins de capsules que celles utilisées par l'ambisonique à l'ordre 1. Les solutions apportées par ce chapitre ont été développées afin d'obtenir un dispositif de captation suffisamment compact pour être utilisé dans un terminal mobile et prenant en compte les contraintes relevées au chapitre II.

Dans notre approche, nous considérons une analyse en temps-fréquence, avec l'hypothèse qu'à chaque trame temporelle, une seule source est présente pour chacune des fréquences [Pulkki, 2006, Berge and Barrett, 2010]. De cette façon, une direction unique est attribuée à cette source. On supposera également que les sources composant la scène sonore sont suffisamment éloignées des capteurs, afin de pouvoir de considérer le champ sonore résultat comme étant composé d'ondes planes.

IV.2 Définitions préalables relatives à l'utilisation de microphones directionnels

Les contraintes de réalisation que nous nous sommes imposées sont les suivantes :

- la localisation des directions des sources sonores dans l'ensemble de directions doit être effectuée avec au maximum 3 microphones,
- les microphones utilisés doivent composer une unité compacte pour pouvoir être utilisés dans un contexte de mobilité.

L'utilisation de microphones omnidirectionnels ne nous paraît donc pas adaptée au cahier de charges, à cause des contraintes liées au nombre de dispositifs nécessaires et à leur espacement. En effet, l'emploi de microphones omnidirectionnels impose en théorie au moins quatre capteurs pour une localisation en trois dimensions. Les microphones doivent être suffisamment éloignés les uns des autres, afin d'estimer une différence de phase dans les basses fréquences. De manière antagoniste, pour éviter le repliement spatial dans les hautes fréquences, ces mêmes capteurs doivent être suffisamment proches les uns des autres, ce qui en augmente considérablement leur nombre.

Par la suite, nous allons proposer plusieurs configurations microphoniques utilisant des capteurs directionnels avec des directivités conventionnelles, en recherchant la meilleure configuration susceptible d'extraire suffisamment d'informations spatiales pour obtenir la localisation des sources.

Les microphones sont des transducteurs électroacoustiques qui peuvent être caractérisés par la façon dont **la traduction acoustique** est effectuée et par leur directivité. Concernant le principe de transduction, on rencontre classiquement les deux types suivants de microphones :

- le microphone électro-dynamique, pour lequel le déplacement d'un bobinage solidaire de la membrane plongée dans un entrefer magnétique engendre un courant à ses bornes,
- le microphone électrostatique, pour lequel la membrane constitue l'une des électrodes d'un condensateur, son déplacement rapproche et éloigne les armatures, faisant ainsi varier la capacité.

La transduction étant un sujet à part entière, nous allons nous focaliser sur le deuxième type de classification : la directivité.

La directivité caractérise le niveau relatif reçu par le microphone en fonction de l'angle d'incidence de l'onde acoustique, et est habituellement représentée par des diagrammes polaires.

Deux directivités "extrêmes" peuvent être détaillées :

- celle du microphone omnidirectionnel : le niveau relatif ne variant pas quelle que soit la direction de la source,
- le microphone bidirectionnel ou à gradient de vitesse : caractérisé par un diagramme dipolaire.

Dans sa conception, le microphone omnidirectionnel peut être décrit par une membrane dont une seule des faces est soumise à la pression acoustique (la deuxième étant "isolée" dans une enceinte close). Le microphone bidirectionnel de son côté a une membrane dont les deux faces sont exposées à la pression acoustique de la même manière, celle-ci s'annulant lorsque l'onde acoustique est perpendiculaire à la membrane.

Toutes les autres directivités usuelles dérivent de ces deux directivités principales et sont obtenues grâce au dosage de l'exposition de la face arrière de la membrane.

Le signal électrique S_n délivré par un microphone n est proportionnel à la pression acoustique exercée sur sa membrane. En fonction de la directivité du microphone, le signal électrique résultant dépend également de la direction de l'onde incidente par rapport au microphone. Nous pouvons donc décrire la directivité d'un microphone n comme une pondération directionnelle de la pression acoustique P_a. Elle est par conséquent décrite par une fonction $M_n(\theta_s, \phi_s)$ dépendant de la position de la source (r_s, θ_s, ϕ_s), en coordonnées sphériques. On a donc

$$S_n = \eta_n M_n P_a, \tag{IV.1}$$

où η_n est la sensibilité du microphone, c'est-à-dire sa capacité de transduction acoustique-électrique définie en V/Pa. Dans la suite du document nous allons définir $\eta = 1$ et pour éviter toute confusion, nous introduirons le signal de pression

$$S_0 \equiv P_a. \tag{IV.2}$$

La relation IV.1 devient

$$S_n = M_n S_0. \tag{IV.3}$$

Dans le cas d'un signal omnidirectionnel, le signal électrique vérifie alors

$$S_n = S_0. \tag{IV.4}$$

Dans le cas d'un microphone bidirectionnel, on a

$$S_n = S_0 \cos(\gamma_n), \tag{IV.5}$$

soit

$$M_n = \cos(\gamma_n),$$

où γ_n est défini par la position de la source par rapport à l'axe du microphone.

Dans le premier cas, M_n définit un monopôle et dans le second un dipôle.

Considérant le cas d'une directivité quelconque, celle-ci peut se décrire comme une somme pondérée du monopôle et du dipôle, suivant

$$M_n = \frac{1}{\delta_n + \zeta_n} \left[\delta_n + \zeta_n \cos(\gamma_n)\right], \tag{IV.6}$$

où δ_n et ζ_n sont les coefficients de pondération. Cette relation est généralisée pour une représentation 3D sous la forme

$$M_n = \frac{1}{\delta_n + \zeta_n} \left[\delta_n + \zeta_n \alpha_n\right], \tag{IV.7}$$

avec

$$\alpha_n = \vec{d_s} \cdot \vec{d_{p_n}}, \tag{IV.8}$$

où $\vec{d_s}$ détermine la direction de la source et $\vec{d_p}$ la direction de pointage du microphone et $\cdot$ le produit scalaire entre ces deux vecteurs.

Ces vecteurs peuvent être exprimés dans une base de coordonnées sphériques $\mathscr{B}_s$ ou de coordonnées cartésiennes $\mathscr{B}_c$ (figure IV.1) sous la forme

$$\vec{d} = \begin{pmatrix} \theta \\ \phi \\ r \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} r\cos\theta\cos\phi \\ r\sin\theta\cos\phi \\ r\ \ \sin\phi \end{pmatrix}_{\mathscr{B}_c}. \tag{IV.9}$$

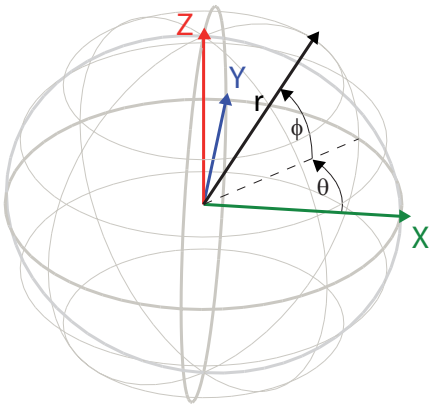


Figure IV.1 : Systèmes de coordonnées cartésiennes ($\vec{x}$,$\vec{y}$,$\vec{z}$) et sphériques ($\vec{\theta}$,$\vec{\phi}$,$\vec{r}$) utilisés dans ce document.

Pour simplifier les notations, $\vec{d}$ est considéré par la suite comme unitaire avec $r = 1$.

La relation (IV.7) peut être reformulée sous la forme

$$M_n = \frac{1}{2}\left[\xi_n + (2 - \xi_n)\alpha_n\right], \tag{IV.10}$$

avec $0 \leq \xi_n \leq 2$, $\xi_n = \delta_n$ et $2 - \xi_n = \zeta_n$ (figure IV.2).

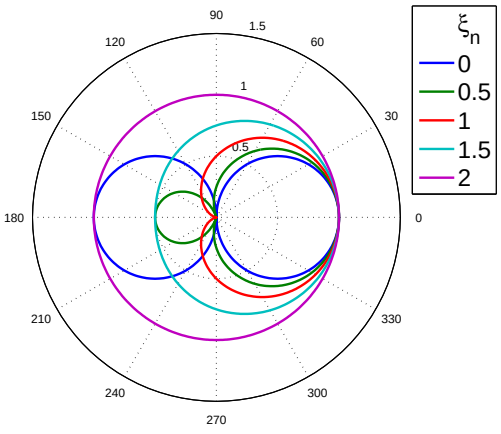


Figure IV.2 : Diagrammes polaires de directivité selon la relation (IV.10) pour des différentes valeurs de ξ_n.

IV.3 Estimation de l'information spatiale basée sur la directivité des capteurs

Par la suite, nous proposons d'utiliser la directivité des capteurs pour localiser la source. En effet, dans l'équation (IV.3) le terme M_n qui définit la directivité est une pondération directionnelle du signal de la source S_o. M_n dépend donc de la position de la source en coordonnées sphériques (θ et ϕ) et est propre au microphone n. Connaissant S_o et les fonctions directionnelles $M_n(\theta, \phi)$, il ne reste alors qu'à résoudre une équation à deux inconnues pour estimer les coordonnées de la direction de la source.

Nous cherchons une configuration microphonique basée sur des directivités décrites par l'équation IV.10, permettant d'extraire le signal de pression S_0 et de fournir suffisamment d'information spatiale pour identifier la direction des sources (une source par trame temps-fréquence).

IV.3.1 Définition de la configuration microphonique à partir de la directivité

IV.3.1.a Localisation dans le plan azimutal

Dans un premier temps, l'étude se limite à localiser la direction de la source dans un plan (problème à 2D). Dans ce cas, les capteurs sont placés dans le plan horizontal $(\vec{x}, \vec{y})$.

Donc, les vecteurs $\vec{d_s}$ et $\vec{d_{p_n}}$, dont les coordonnées sont données par la relation IV.9 (avec $\phi = 0$), deviennent,

$$\vec{d} = \begin{pmatrix} \theta \\ 0 \\ 1 \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix}_{\mathscr{B}_c} . \tag{IV.11}$$

IV.3.1.b Microphone unique

Le signal de pression S_o peut être obtenu directement à l'aide d'un microphone respectant la relation (IV.10), avec $\xi = 2$ (microphone omnidirectionnel). Dans ce cas, aucune information directionnelle ne peut être extraite, car la fonction M_1 ne dépend pas de la position de la source. Dans le cas contraire, c'est-à-dire avec l'utilisation d'un microphone

directionnel unique respectant la relation (IV.10) avec $\xi \neq 2$, la fonction M_1 dépend de la position de la source, mais aucune référence ne permet de extraire la direction, car l'information du signal de pression S_o est manquante.

Cette information complémentaire ne peut donc être trouvée qu'en utilisant au moins deux capteurs.

Nous définissons par la suite une fonction $\mathcal{N}(\theta)$ modélisant la relation qui permet d'estimer la direction θ. Celle-ci est obtenue à partir d'une relation linéaire des signaux S_n composant le réseau microphonique de N capteurs, notamment à partir des directivités M_n de ces derniers.

Dans une configuration idéale, la précision sur la localisation de θ doit être identique quelle que soit la valeur de θ. La quantité $|\mathrm{d}\mathcal{N}/\mathrm{d}\theta|$ doit donc être proche d'une constante non nulle. Comme les fonctions qui interviennent dans $\mathcal{N}(\theta)$ sont périodiques, nous cherchons une configuration microphonique nous permettant de nous rapprocher d'une fonction $|\mathrm{d}\mathcal{N}/\mathrm{d}\theta|$ constante non nulle par morceaux.

IV.3.1.c Couple formé par un microphone omnidirectionnel et un microphone directionnel

Nous étudions d'abord l'utilisation de deux microphones coïncidents dont le premier est omnidirectionnel ($M_1 = 1$) et le second est un microphone comportant une directivité M_2 (IV.10), avec $\xi_2 \neq 2$ (figure IV.2). Cette configuration permet d'obtenir directement les informations recherchées, à savoir, le signal de pression S_o avec le microphone omnidirectionnel et la direction en combinant les informations S_1 et S_2.

Le microphone directionnel pointe dans la direction :

$$\vec{d}_{p_2} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}_{\mathscr{B}_c} . \tag{IV.12}$$

On a alors

$$\alpha_2 = \cos\theta \tag{IV.13}$$

et l'équation (IV.10) devient

$$M_2 = \frac{1}{2}\left[\xi + (2-\xi)\cos\theta\right]. \tag{IV.14}$$

Pour considérer deux cas réalistes, examinons le cas de deux directivités M_2 avec $\xi = 0$ et $\xi = 1$. Ces deux valeurs correspondent respectivement à un microphone bidirectionnel et à un microphone cardioïde (figure IV.3).

La relation $\mathcal{N}(\theta)$ associée à ces deux s'écrit alors

$$\mathcal{N}(\theta) = \frac{S_2}{S_1} = \frac{M_2 S_0}{M_1 S_0} = M_2, \tag{IV.15}$$

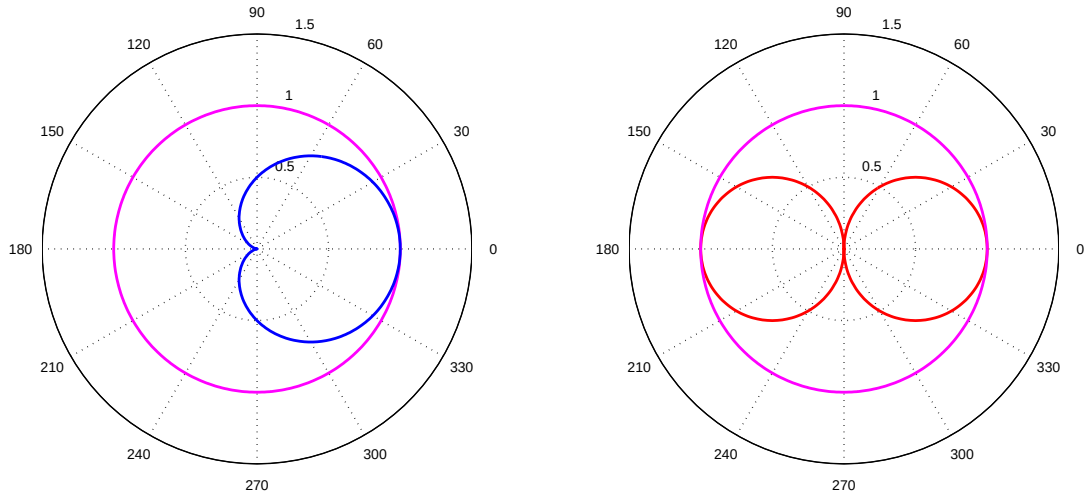
car $M_1 = 1$.

La fonction $\mathcal{N}(\theta)$ s'exprime donc respectivement comme

$$\mathcal{N}(\theta) = \begin{cases} \cos\theta & \text{pour} \quad \xi = 0 \\ \frac{1}{2}[1 + \cos\theta] & \text{pour} \quad \xi = 1 \end{cases}. \tag{IV.16}$$

En étudiant la variation de la directivité $|\mathrm{d}\mathcal{N}/\mathrm{d}\theta|$ en fonction de la direction θ (figure IV.4), on constate, d'une part, que la dynamique fournie par le microphone bidirectionnel est supérieure à celle du microphone cardioïde. D'autre part, les courbes $|\mathrm{d}\mathcal{N}/\mathrm{d}\theta|$ ont des allures sensiblement identiques. Elles atteignent en particulier des valeur nulles lorsque la source se trouve à des directions $\beta\pi$ avec $\beta \in \mathbb{N}$.

L'utilisation d'un couple composé d'un microphone omnidirectionnel et d'un microphone bidirectionnel tend à mieux estimer les directions se trouvant à proximité de $\pi/2 + \beta\pi$, car le gradient dynamique $|\mathrm{d}\mathcal{N}/\mathrm{d}\theta|$ est plus élevé dans ces directions. Un microphone bidirectionnel semble donc être le mieux adapté de par sa dynamique directionnelle. Paradoxalement, pour ce microphone, les directions où le gradient dynamique est le plus fort, correspondent à ses points "sourds" ($M(\theta) = 0$). D'un point de vue pratique, le rapport signal à bruit dans ces zones est très faible. Il introduit donc une erreur dans l'estimation de la position lorsque la source s'y trouve.



(a) Omnidirectionnel et cardioïde . **(b)** Omnidirectionnel et bidirectionnel .

Figure IV.3 : Diagrammes polaires de directivité du couple composé d'un microphone omnidirectionnel et d'un microphone directionnel.

IV.3.1.d Couple directionnel

La seconde configuration étudiée correspond à deux microphones directionnels permettant, d'une part, d'extraire le signal de pression, et d'autre part, d'estimer l'information directionnelle.

Par la suite, nous considérons que le premier de deux microphones pointe vers l'axe $\vec{x}$ (IV.12) et que le second capteur pointe dans une direction quelconque θ_{p2}. Ainsi, on a

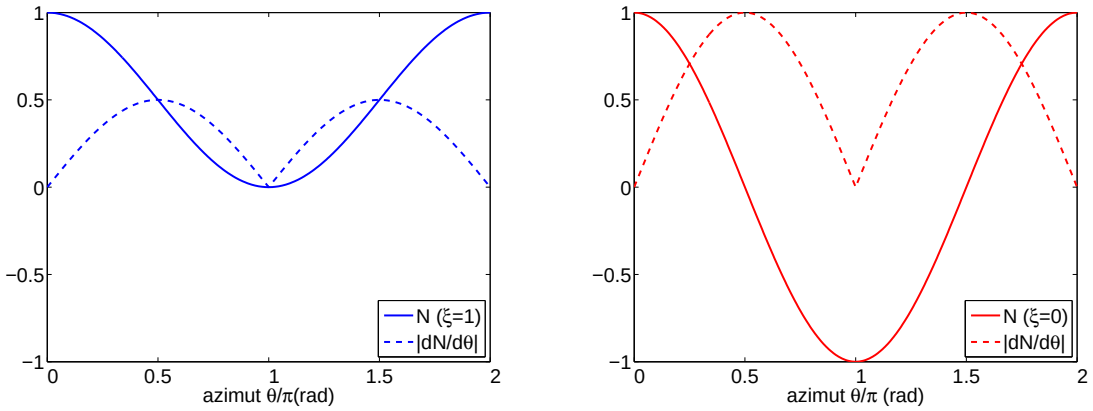
$$\vec{d_{p2}} = \begin{pmatrix} \theta_{p2} \\ 0 \\ 1 \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} \cos\theta_{p2} \\ \sin\theta_{p2} \\ 0 \end{pmatrix}_{\mathscr{B}_c} . \tag{IV.17}$$

Les fonctions de directivité M_1 et M_2 associées à l'équation (IV.10) correspondent alors à l'équation (IV.14) avec $\xi_1 = \xi_2 = \xi$, d'où

$$M_n = \frac{1}{2}\left[\xi + (2-\xi)\alpha_n(\theta)\right], \tag{IV.18}$$

avec

$$\alpha_1 = \cos\theta \tag{IV.19}$$



(a) **(b)**

Figure IV.4 : Fonctions directionnelles $\mathcal{N}$ (traits continus) et leur variation $\frac{\mathrm{d}\mathcal{N}}{\mathrm{d}\theta}$ (traits pontillés) pour : (a) un microphone cardioïde et (b) un microphone bidirectionnel, en fonction de la direction de la source.

et

$$\alpha_2 = \cos\theta\cos\theta_{p2} + \sin\theta\sin\theta_{p2}. \tag{IV.20}$$

Deux situations particulières seront analysées par la suite. La première où $\theta_{p2} = \pi/2$ et $\xi = 0$ (couple bidirectionnel perpendiculaire) et la seconde où $\theta_{p2} = \pi$ et $\xi = 1$ (couple cardioïde) (figure IV.5).

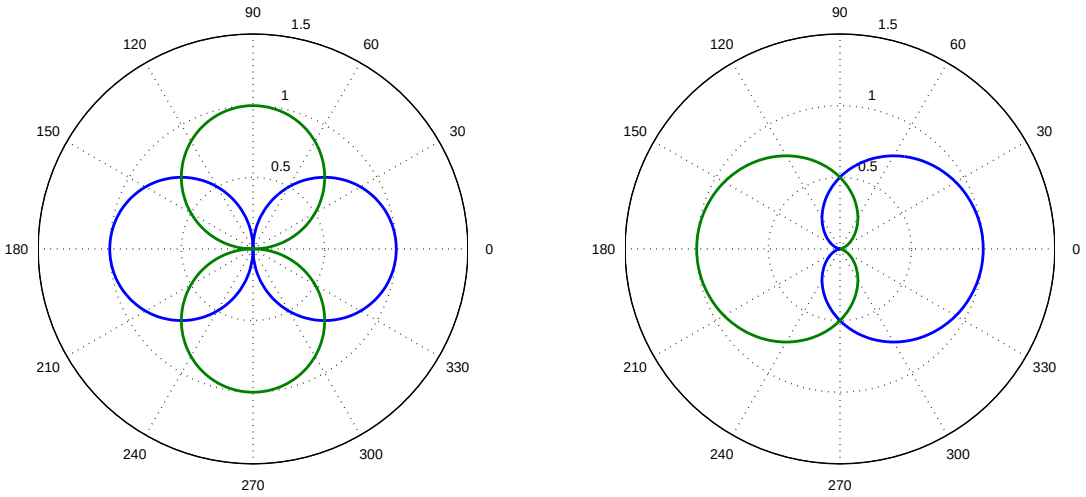
Couple bidirectionnel Nous considérons tout d'abord un **couple bidirectionnel** $(\xi_1 = \xi_2 = 0)$ dont le deuxième microphone est perpendiculaire au premier ($\theta_{p1} = 0$ et $\theta_{p2} = \pi/2$) (figure IV.5).

Les directivités M_1 et M_2 des microphones 1 et 2 extraites de la relation (IV.10) deviennent

$$\begin{cases} M_1(\theta) = \cos\theta \\ M_2(\theta) = \sin\theta \end{cases}. \tag{IV.21}$$

Dans cette configuration, il est possible d'accéder au signal de pression en utilisant la relation

$$S_o^2 = S_1^2 + S_2^2, \tag{IV.22}$$



(a) Couple bidirectionnel **(b)** Couple cardioïde

Figure IV.5 : Diagrammes polaires de directivité M_1 et M_2 (a) du couple bidirectionnel et (b) du couple cardioïde.

où S_1 et S_2 correspondent respectivement aux signaux délivrés par les microphones 1 et 2. L'énergie acoustique induite par la source à l'emplacement des microphones est donc égale à la somme des énergies délivrées par les capsules.

Cette configuration microphonique permet d'utiliser les relations (IV.3) et (IV.14) pour estimer la direction de la source grâce aux fonctions $\mathcal{N}$

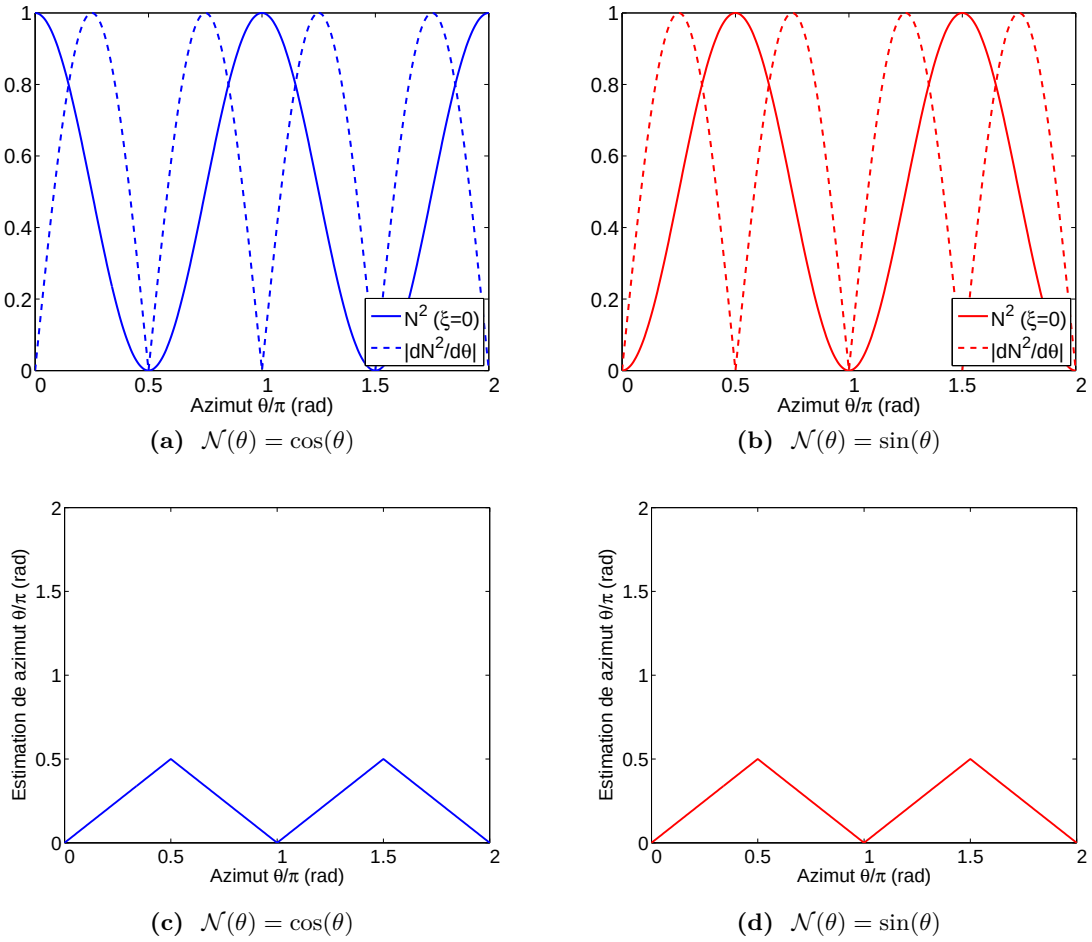
$$\mathcal{N}(\theta) = \cos(\theta) = \sqrt{\frac{S_1^2}{S_1^2 + S_2^2}} \tag{IV.23a}$$

ou

$$\mathcal{N}(\theta) = \sin(\theta) = \sqrt{\frac{S_2^2}{S_1^2 + S_2^2}}. \tag{IV.23b}$$

L'utilisation de cette expression est illustrée en figures IV.6a et IV.6b. On y observe que la dynamique angulaire est réduite par rapport à celle obtenue en figure IV.4b, du fait qu'on utilise la valeur quadratique des signaux. La variation la plus faible est obtenue pour des directions $\beta\pi/2$. Les figures IV.6c et IV.6d affichent le "repliement spatial" obtenu avec l'utilisation des relations (IV.23). En effet, toutes les sources sont localisées dans le premier quart du cercle $[0, \pi/2]$, ce qui nous conduit à écarter cette solution.

Cependant, avec cette même configuration microphonique, il est possible d'utiliser la



(a) $\mathcal{N}(\theta) = \cos(\theta)$ **(b)** $\mathcal{N}(\theta) = \sin(\theta)$

(c) $\mathcal{N}(\theta) = \cos(\theta)$ **(d)** $\mathcal{N}(\theta) = \sin(\theta)$

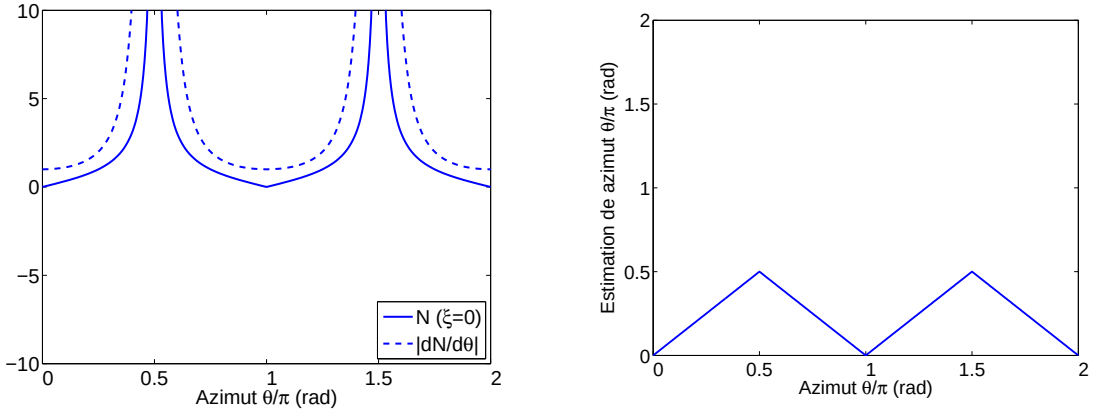
Figure IV.6 : Énergie délivrée par des microphones bidirectionnels $\mathcal{N}^2(\theta)$ (lignes continues) et leur variation angulaire énergétique $\left|\mathrm{d}\mathcal{N}^2/\mathrm{d}\theta\right|$ (lignes pointillées) en fonction de la direction de la source pour des microphones pointant vers $\vec{x}$ (a) et vers $\vec{y}$ (b). Estimation de la direction correspondante obtenue grâce à la relation (c) (IV.23a) et (d) (IV.23b).

relation $\mathcal{N}$,

$$\mathcal{N}(\theta) = \tan(\theta) = \frac{S_2}{S_1}. \tag{IV.24}$$

Cette expression permet en effet d'obtenir une forte variation pour les directions proches de $\pi/2 + \beta\pi$ (figure IV.7).

L'utilisation de cette configuration microphonique est écartée, car même si elle permet une localisation angulaire correcte en utilisant les relations IV.23, le signal de pression S_o ne peut pas être estimé (seule sa valeur quadratique peut l'être).



(a) **(b)**

Figure IV.7 : (a) Variation angulaire en fonction de la direction θ de la source pour une paire de microphones bidirectionnels perpendiculaires. (b) Estimation de la direction correspondante obtenue grâce à la relation (IV.24) $\mathcal{N}(\theta) = \tan(\theta)$.

Couple cardioïde Avec l'utilisation d'un couple cardioïde ($\xi = 1$), nous envisageons la configuration pour laquelle le premier microphone pointe vers $\vec{x}$ et le second vers $-\vec{x}$.

Dans cette configuration, les directivités du premier et du second microphone sont respectivement

$$\begin{cases} M_1(\theta) = & \frac{1}{2}\left[1 + \cos\theta\right] \\ M_2(\theta) = & \frac{1}{2}\left[1 - \cos\theta\right] \end{cases}. \tag{IV.25}$$

Le signal de pression S_0 est directement accessible par la relation

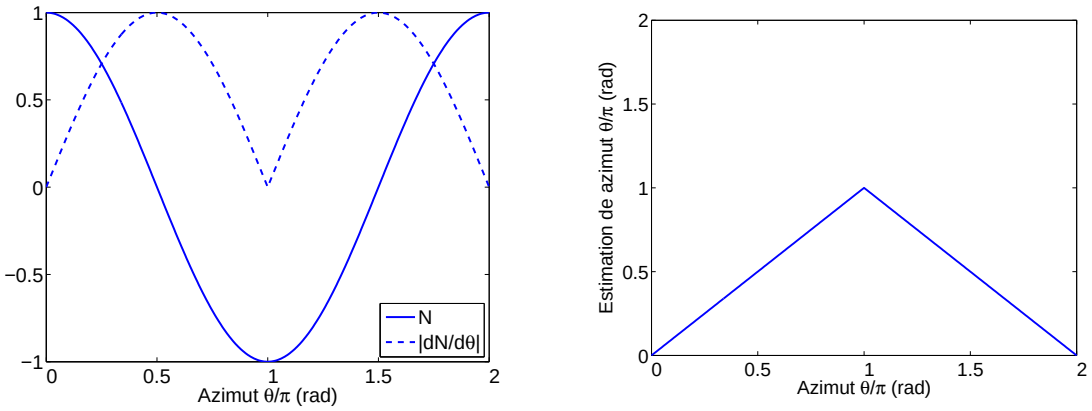
$$S_0 = S_1 + S_2 \tag{IV.26}$$

et la direction de la source peut être obtenue selon la relation

$$\mathcal{N}(\theta) = \cos(\theta) = \frac{S_1 - S_2}{S_1 + S_2}. \tag{IV.27}$$

Dans cette configuration, $\mathcal{N}(\theta)$ est équivalent (figure IV.8a) à l'équation (IV.16) dans la configuration "couple composé d'un microphone omnidirectionnel et d'un microphone bidirectionnel".

Cependant, à la différence de la relation (IV.16), $\mathcal{N}(\theta)$ ne dépend pas de la directivité d'un seul des capteurs. Dans ce cas, $\mathcal{N} = 0$ est atteint grâce à l'interférence destructive provoquée par la soustraction des signaux issus de deux capteurs. Les passages par



(a) **(b)**

Figure IV.8 : (a) Variation angulaire en fonction de la direction θ de la source pour une paire de microphones cardioïdes à 180°. (b) Estimation de la direction correspondante obtenue grâce à la relation (IV.27).

zéro ne correspondent pas aux points "sourds" des capteurs et ces valeurs ne sont pas perturbés par la chaîne électroacoustique, comme c'était le cas de la configuration mixte (microphone omnidirectionnel associé à un microphone bidirectionnel).

Comme illustré en figure (IV.8a), un repliement spatial est obtenu de par la fonction $\cos^{-1}$ de l'équation (IV.27).

Cette configuration microphonique est retenue par la suite, car elle présente un avantage considérable de par sa simplicité de mise en œuvre. Néanmoins, deux problèmes restent à résoudre : la résolution de l'ambiguïté de localisation et son utilisation pour un usage en trois dimensions.

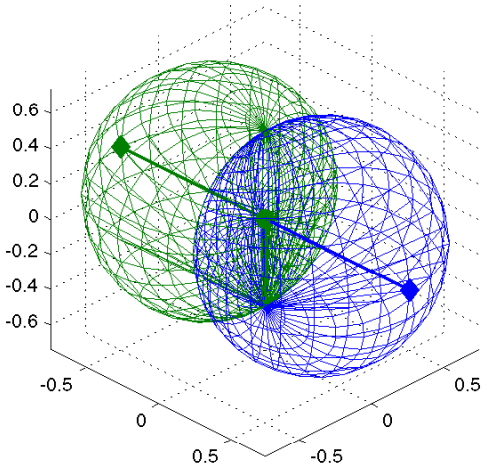


Figure IV.9 : Directivité microphonique M_1 et M_2 dans l'espace 3D d'un couple cardioïde coïncidant dont les capsules pointent respectivement vers $\vec{x}$ et $-\vec{x}$.

IV.3.1.e Localisation dans l'espace 3D

Considérant la configuration microphonique que nous avons retenue dans le paragraphe précédent (couple cardioïde dont les capsules pointent respectivement vers $\vec{x}$ et vers $-\vec{x}$), les performances de cette configuration sont analysées dans l'espace 3D (figure IV.9).

Dans l'espace 3D, ces capteurs ont une directivité M_1 et M_2 qui correspond à celle décrite dans l'équation (IV.10).

Elle est calculée à l'aide des vecteurs $\vec{d_s}$ correspondant à la relation (IV.9) et les vecteurs de pointage $\vec{d_{p1}}$ et $\vec{d_{p2}}$ associés deviennent

$$\vec{d_{p1}} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}_{\mathscr{B}_c} \tag{IV.28}$$

et

$$\vec{d_{p2}} = \begin{pmatrix} \pi \\ 0 \\ 1 \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}_{\mathscr{B}_c} . \tag{IV.29}$$

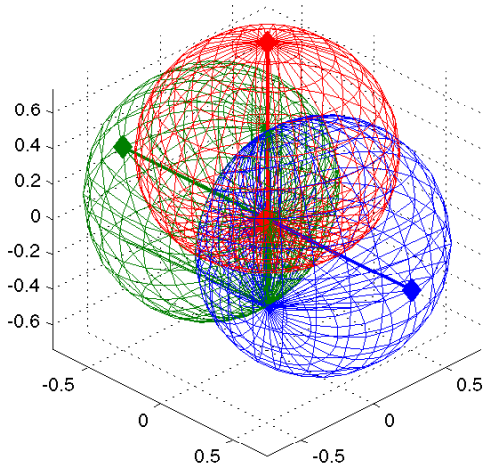


Figure IV.10 : Représentation des directivités microphoniques M_1, M_2 et M_3 dans l'espace 3D d'un triplet cardioïde coïncidant dont les capsules pointent respectivement vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$.

Les directivités correspondantes s'écrivent

$$\begin{cases} M_1(\theta,\phi) &= \frac{1}{2}(1+\cos\theta\cos\phi) \\ M_2(\theta,\phi) &= \frac{1}{2}(1-\cos\theta\cos\phi) \end{cases}. \tag{IV.30}$$

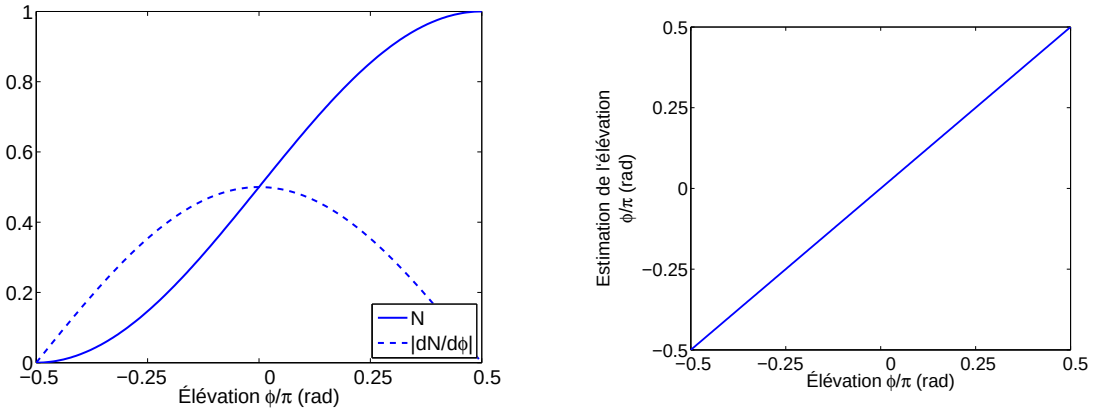
Avec la prise en compte de la troisième dimension et de l'élévation de la source, le signal de pression se calcule de la même manière que pour la version 2D (équation (IV.26)) et la fonction $\mathcal{N}(\theta)$ associée est

$$\mathcal{N}(\theta,\phi) = \cos\theta = \frac{S_1 - S_2}{\cos\phi(S_1+S_2)}. \tag{IV.31}$$

Dans cette version, on remarque que l'expression de l'azimut est dépendante de l'élévation ϕ. Cette dernière ne peut pas être obtenue avec les deux capteurs retenus initialement. L'ajout d'un troisième capteur est nécessaire afin d'estimer cette valeur.

Nous considérons donc l'utilisation d'un troisième capteur cardioïde pointant vers $\vec{z}$ (figure IV.10). Le vecteur de pointage le caractérisant est donné par

$$\vec{d_{p3}} = \begin{pmatrix} 0 \\ \pi/2 \\ 1 \end{pmatrix}_{\mathscr{B}_s} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}_{\mathscr{B}_c} \tag{IV.32}$$



(a) **(b)**

Figure IV.11 : (a) Variation angulaire en fonction de l'élévation de la source pour un microphone pointant vers $\vec{z}$ et (b) estimation de la direction correspondante obtenue grâce à la relation (IV.34).

et sa directivité par

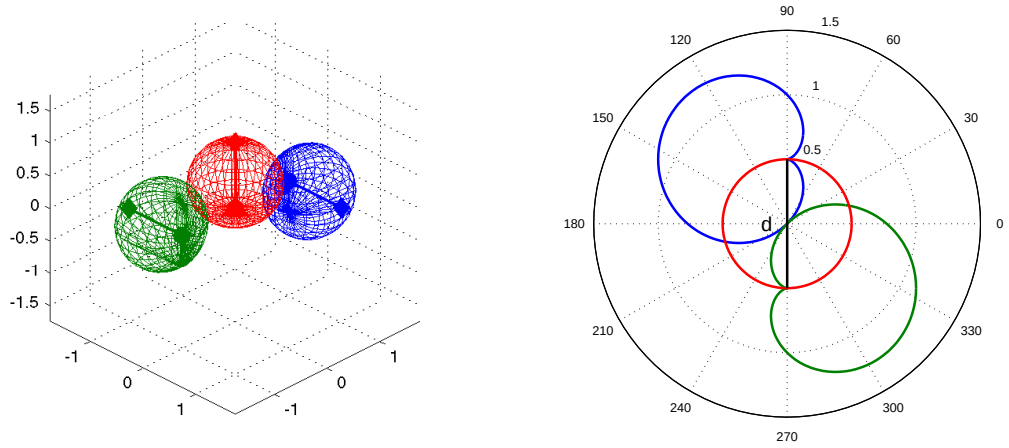
$$M_3(\theta, \phi) = \frac{1}{2}(1 + \sin\phi). \tag{IV.33}$$

D'après cette relation, $\mathcal{N}(\phi)$ s'écrit

$$\mathcal{N}(\phi) = \sin\phi = \frac{2S_3}{S_1 + S_2} - 1. \tag{IV.34}$$

Comme l'illustre la figure IV.11, la variation de $\mathcal{N}(\phi)$ possède des minima à $\pm\pi/2$. Il est possible de profiter de cette caractéristique en plaçant le corps du microphone ainsi que son dispositif de maintien à $-\pi/2$. Lorsque la source se trouve près de $\pi/2$, l'acuité perceptive est dégradée (MAA élevée) et une erreur sur cette valeur n'est pas perçue par l'auditeur (I.2.4).

Dorénavant, nous pouvons donc considérer que la configuration microphonique optimale pour la localisation des sources est celle à base de trois capsules microphoniques cardioïdes, pointant respectivement vers $\vec{x}$, $-\vec{x}$, et $\vec{z}$. Néanmoins, l'ambigüité de la localisation en azimut reste à résoudre.



(a) Vue 3D. **(b)** Coupe dans le plan horizontal.

Figure IV.12 : Configuration et directivités microphoniques M_1, M_2 et M_3 permettant l'exploitation du retard entre les deux capteurs (1 et 2) pointant dans le plan horizontal.

IV.3.2 Résolution de l'ambiguïté sur l'estimation de l'azimut en exploitant le retard entre les capteurs

Tel qu'annoncé en IV.3.1.e, une erreur entache la localisation de l'azimut de la source. Pour la corriger sans ajouter de capteurs supplémentaires, nous proposons (figure IV.12) d'utiliser le retard entre les signaux issus de deux microphones pointant dans le plan horizontal (capteurs 1 et 2).

En effet, comme il est impossible dans une configuration réelle de placer ces deux capteurs au même point, nous profitons de ce décalage spatial pour utiliser le retard entre ces deux capteurs de manière à lever l'ambiguïté.

Considérant qu'un microphone n se trouve à la position définie par le vecteur

$$\vec{E_n} = \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}_{\mathscr{B}_c}, \tag{IV.35}$$

le signal S_n issu de ce microphone et défini par l'équation (IV.3) s'écrit

$$S_n(t,\theta,\phi) = M_n(\theta,\phi)S_o(t-\tau_n), \tag{IV.36}$$

où τ_n est le retard induit par la distance entre le microphone n et l'origine du repère. Considérant que la source émet des ondes planes, ce retard vaut

$$\tau_n = \frac{1}{c}\,\vec{d_s} \cdot \vec{E_n}. \tag{IV.37}$$

Dans le domaine fréquentiel, $S_n(t)$ devient

$$\begin{aligned} S_n(\omega,\theta,\phi) &= \mathrm{TF}\left[S_n(t,\theta,\phi)\right] \\ &= M(\theta,\phi)S_o(\omega)e^{-j\omega\tau_n}, \end{aligned} \tag{IV.38}$$

où $\mathrm{TF}\left[g(t)\right]$ représente la transformée de Fourier de $g(t)$. Considérant que

$$S_o(\omega) = |S_o(\omega)|e^{j\angle S_o(\omega)}, \tag{IV.39}$$

la relation (IV.38) peut être développée suivant

$$S_n(\omega,\theta,\phi) = M(\theta,\phi)|S_o(\omega)|e^{-j(\angle S_o(\omega)-\omega\tau_n)}. \tag{IV.40}$$

Par identification, on a donc

$$\angle S_n(\omega,\theta,\phi) = \angle S_o(\omega) - \omega\tau. \tag{IV.41}$$

Par conséquent, on en déduit que

$$\angle S_1(\omega,\theta,\phi) - \angle S_2(\omega,\theta,\phi) = \omega(\tau_1 - \tau_2). \tag{IV.42}$$

Le retard entre les microphones 1 et 2, $\tau_{12} = \tau_1 - \tau_2$, s'écrit finalement

$$\tau_{12} = -\frac{1}{\omega}(\angle S_1 - \angle S_2). \tag{IV.43}$$

L'ambigüité peut donc être résolue en utilisant le signe de cette expression dans (IV.31) et elle devient alors

$$\theta = \frac{\tau_{12}}{|\tau_{12}|} \cos^{-1}\left[\frac{S_1 - S_2}{\cos\phi\,(S_1 + S_2)}\right]. \tag{IV.44}$$

Comme un retard est introduit, un filtrage en peigne apparaît pour les fréquences dont la longueur d'onde est proche de l'écart inter-microphonique. En prenant en compte cette considération, l'écart inter-microphonique doit être choisi en fonction de la bande de fréquences où l'on souhaite effectuer la localisation des sources.

IV.3.3 Performances de localisation

Les performances de la configuration microphonique et de l'algorithme de localisation associé ont été évaluées à partir de sources synthétiques. L'interaction des sources avec les microphones a été simulée grâce aux relations (IV.30), (IV.33) et (IV.36).

A titre d'illustration, nous utilisons une source large bande tournant autour du dispositif microphonique, à des élévations variant par palier de 20°, partant de l'hémisphère sud vers le zénith. La trajectoire de cette source est illustrée en figure IV.14.

IV.3.3.a Critères d'évaluation

Afin d'évaluer les performances de localisation, des indicateurs permettant de mesurer l'erreur d'estimation sont définis.

Dans l'approche directe, nous définissons les erreurs $E_\theta(t,f)$, $E_\phi(t,f)$ et $E_t(t,f)$ comme une mesure de la distance angulaire entre direction cible $\vec{d}_s(t,f)$ et direction estimée $\vec{d}_{est}(t,f)$ par l'algorithme de localisation. Ces valeurs sont calculées grâce au produit scalaire

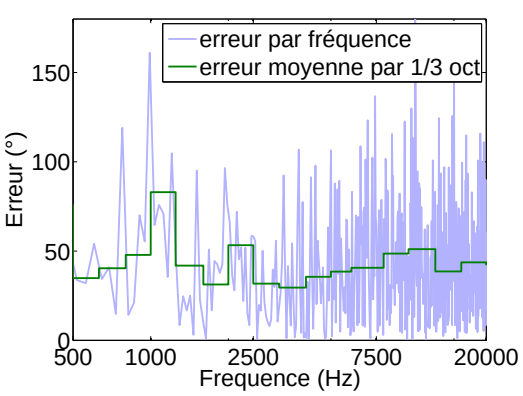
$$E(t,f) = \cos^{-1}\left(\frac{\vec{V}_{cib}(t,f)\cdot\vec{V}_{est}(t,f)}{\|\vec{V}_{cib}(t,f)\|\|\vec{V}_{est}(t,f)\|}\right), \tag{IV.45}$$

où

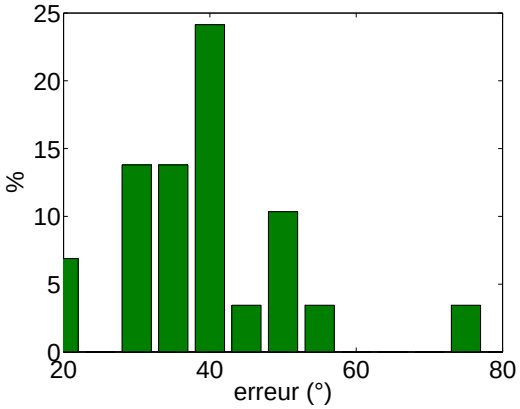
$$\vec{d}_s(t,f) = \begin{pmatrix} \theta_s(t,f) \\ \phi_s(t,f) \\ 1 \end{pmatrix}_{\mathscr{B}_s} \quad \text{et} \quad \vec{d}_{est} = \begin{pmatrix} \theta_{est}(t,f) \\ \phi_{est}(t,f) \\ 1 \end{pmatrix}_{\mathscr{B}_s} . \tag{IV.46}$$

Pour l'évaluation de l'erreur dans une seule des coordonnées $E_\theta(t,f)$ ou $E_\phi(t,f)$, les composantes de la coordonnée non évaluée sont respectivement fixées à 0. C'est-à-dire $\phi_{cib}(t,f) = 0$ et $\phi_{est}(t,f) = 0$ pour le calcul de $E_\theta(t,f)$ et $\theta_{cib}(t,f) = 0$ et $\theta_{est}(t,f) = 0$ pour le calcul de $E_\phi(t,f)$.

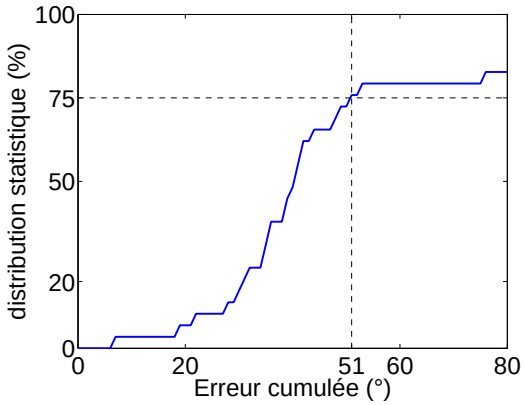
Cette erreur est exprimée en degrés et sa valeur est de $0°$ lorsque l'estimation est parfaite et de $180°$ dans le pire des cas. Compte tenu que l'analyse est effectuée conjointement en temps et en fréquence, un indice est proposé, permettant à chaque instant d'extraire une valeur unique sur l'ensemble du spectre. Il s'agit de l'indicateur statistique $E_{75}(t)$, qui est l'erreur obtenue par au moins 75% du spectre en tiers d'octaves (figure IV.13).



(a) Distribution spectrale de l'erreur.

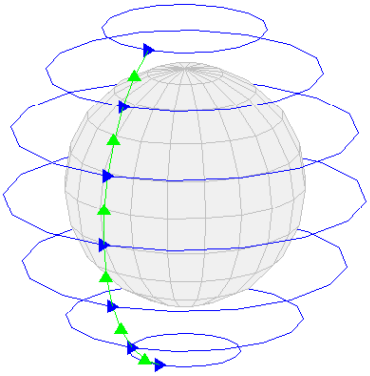


(b) Histogramme de la distribution statistique de l'erreur en tiers d'otaves.

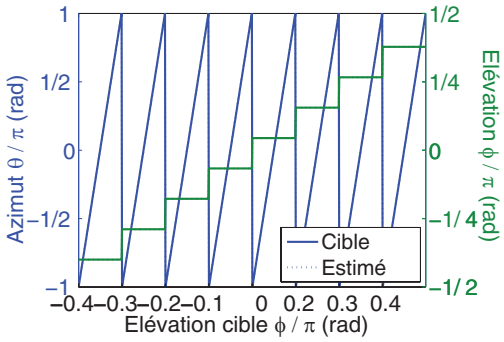


(c) Estimation du E_{75}.

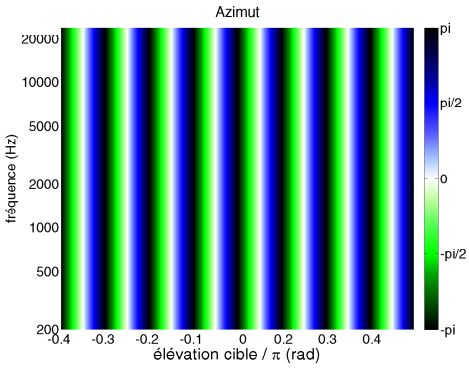
Figure IV.13 : Étapes pour le calcul du E_{75}.



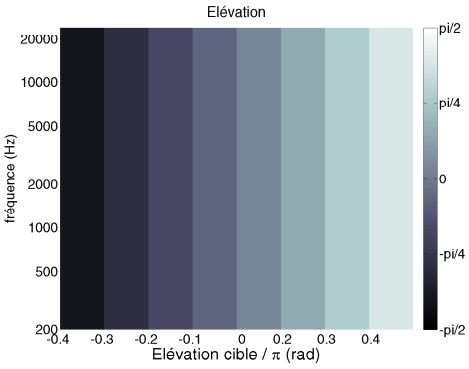
(a) Trajectoire.



(b) Trajectoire à 1 kHz.



(c) Représentation conjointe position-fréquence du déplacement de la source en azimut.



(d) Représentation conjointe position-fréquence du déplacement de la source en élévation.

Figure IV.14 : Trajectoire de la source sonore virtuelle : (a) et (b) Le déplacement en azimut est illustré par les flèches et la courbe bleues et en élévation par les flèches et la courbe vertes. (a) illustre le déplacement dans l'espace 3D. (b) illustre la trajectoire en azimut et en élévation de la source pour une fréquence de 1 kHz. Les figures (c) et (d) sont la représentation conjointe position-fréquence du déplacement de la source en azimut (c) et élévation (d). Les positions estimées sont représentées dans une échelle de couleurs pour les fréquences (en ordonnée) en fonction de la position de la source en abscisse (élévation). En (c) l'azimut est représenté dans un dégradé de verts lorsque $\theta \in [-\pi, 0]$ et de bleus lorsque $\theta \in [0, \pi]$. Les couleurs foncées indiquent la proximité à π et les couleurs claires indiquent la proximité à 0. En (d) l'élévation de la source est représentée en échelle de gris où les tons foncés indiquent que la source est proche du pôle sud ($\phi \sim -\pi/2$) et les tons clairs du pôle nord ($\phi \sim \pi/2$).

IV.3.3.b Analyse des résultats

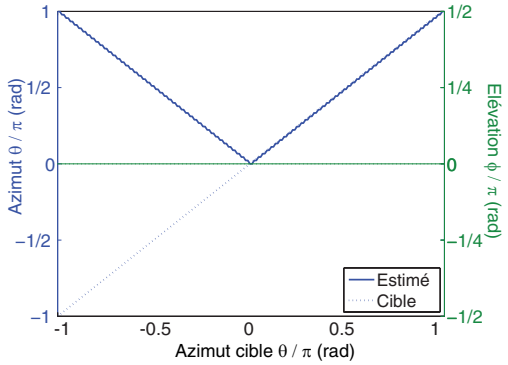
Localisation avec deux microphones Le système microphonique coïncidant, composé de deux capsules microphoniques, a été testé en considérant que la **source sonore est placée dans le espace 2D** $(\vec{x}, \vec{y})$.

La figure IV.15 illustre l'ambiguïté de la localisation. En effet, la source est toujours identifiée sur la plage $[0, \pi]$. Cette erreur peut être vue, soit comme une confusion hémisphérique avant-arrière lorsque l'axe $\vec{y}$ pointe vers l'avant, soit comme une confusion gauche-droite lorsque la partie frontale de la scène est pointée par $\vec{x}$. Nous allons utiliser cette dernière représentation dans la suite du document.

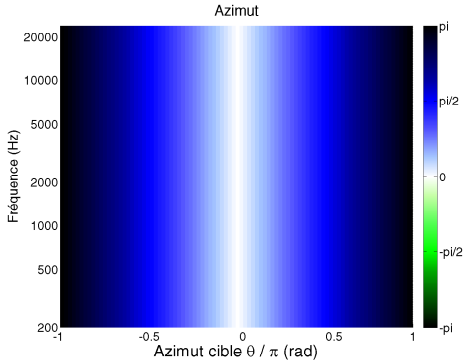
Lorsque les microphones sont écartés l'un de l'autre sur l'axe $\vec{y}$, d'une distance d, l'ambiguïté de localisation disparaît complètement, comme l'illustre la figure IV.16.

Si un tel dispositif est utilisé pour la localisation d'une **source dans l'espace 3D**, l'erreur de localisation de l'azimut croît lorsque la source s'éloigne de l'équateur et est symétrique par rapport à cette ligne (figure IV.17d). Cette erreur apparaît car l'équation (IV.44) est dépendante de ϕ dont la valeur est ici négligée.

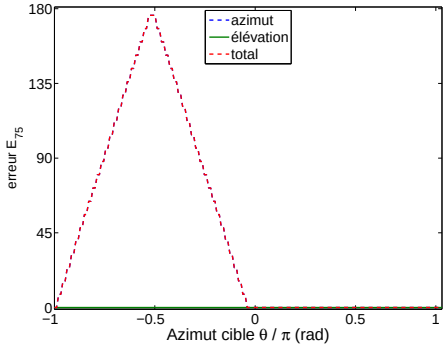
En présence d'une source perturbatrice se trouvant à la direction (0,0) (figure IV.18), les directions ne sont estimées correctement que lorsque le rapport signal-à-bruit (RSB) est supérieur à 15 dB. L'indicateur E_{75} est alors inférieur à 15° lorsque la source se trouve à proximité du plan horizontal. Pour les RSB inférieurs, l'azimut est correctement estimé uniquement lorsque la source perturbatrice et la source utile se trouvent sur le même azimut.



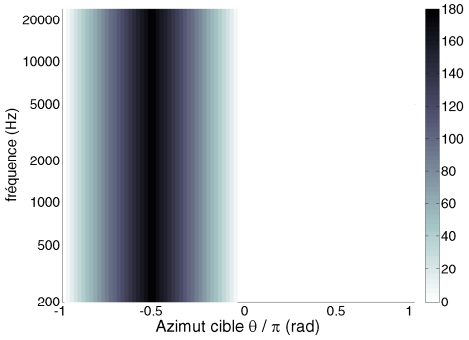
(a) Azimut θ (bleu), Élévation ϕ (vert).



(b) Azimut θ.

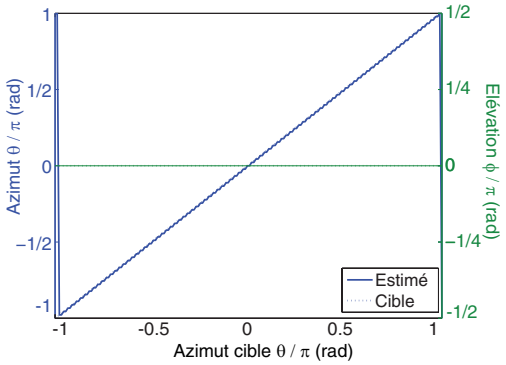


(c) Erreur E_{75} (°).

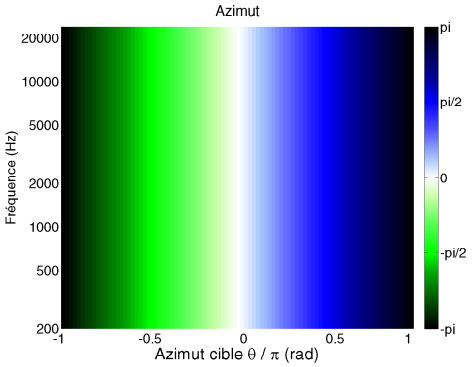


(d) Erreur E_t (°).

Figure IV.15 : Localisation d'une source large bande tournant sur le plan horizontal autour du couple microphonique coïncidant composé de deux capsules cardioïdes. (a) Estimation de l'azimut (en bleu) et de l'élévation (en vert) à une fréquence $f = 1\ kHz$ en fonction de la position cible de la source. (b) Estimation de l'azimut en fonction de la fréquence et la position cible de la source. (c) Erreur E_{75} évalué en fonction de la position cible de la source. (d) Erreur E_t évalué en fonction de la fréquence et la position cible de la source.

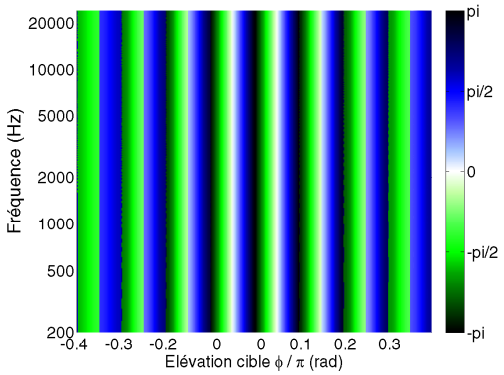


(a) Estimation de la localisation de la source en azimut θ (bleu) et élévation $\phi = 0$ (vert).

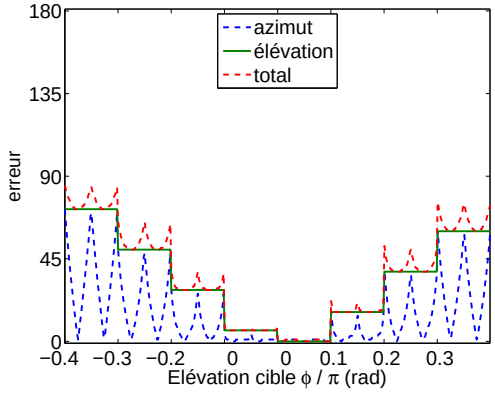


(b) Représentation conjointe position-fréquence de l'estimation de la localisation de la source en azimut θ.

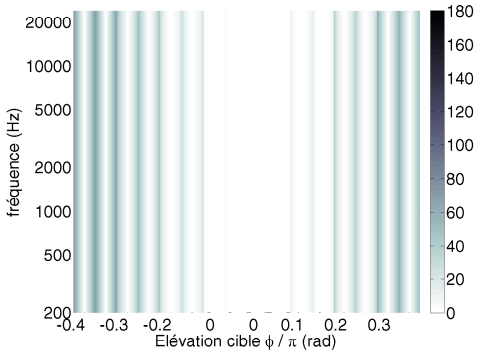
Figure IV.16 : Localisation d'une source large bande tournant sur le plan horizontal autour du couple microphonique cardioïde composé de deux capsules cardioïdes écartées d'une distance d selon l'axe $\vec{y}$. (a) Estimation de de azimut (en bleu) et de l'élévation (en vert) à une fréquence $f = 1\ kHz$ en fonction de la position cible de la source. (b) Estimation de l'azimut en fonction de la fréquence et la position cible de la source.



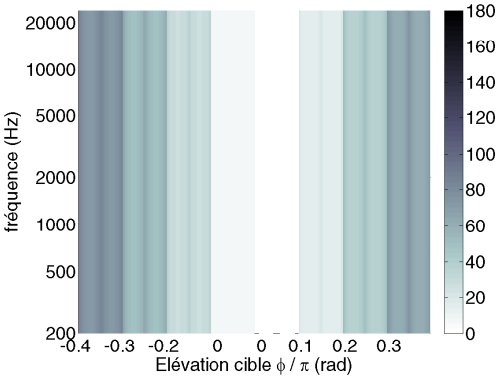
(a) Estimation de azimut θ dans une représentation conjointe position-fréquence.



(b) Erreur E_{75} (°).



(c) Erreur de localisation sur l'azimut E_θ dans une représentation conjointe position-fréquence.

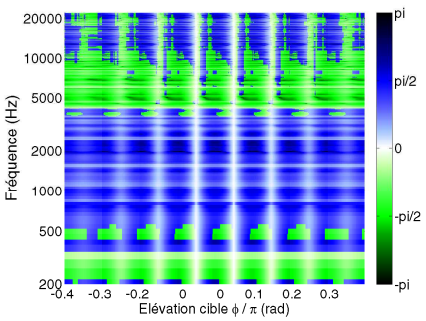


(d) Représentation conjointe position-fréquence de l'erreur de localisation E_t (°).

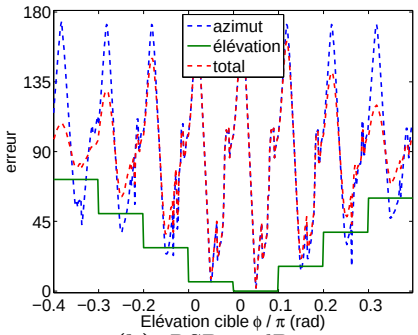
Figure IV.17 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14.(a) Estimation de l'azimut en fonction de la fréquence et la position cible de la source. (b) Erreur d'estimation E_{75} en fonction de la position cible de la source. (c) Erreur d'estimation de l'azimut E_θ en fonction de la fréquence et la position cible de la source. (d) Erreur d'estimation E_t en fonction de la fréquence et la position cible de la source. Résultats obtenus avec l'utilisation d'une antenne microphonique cardioïde composée de 2 capteurs pointant vers $\vec{x}$, $-\vec{x}$. Les capsules sont écartées de 2 cm selon l'axe $\vec{y}$.

Azimut θ :

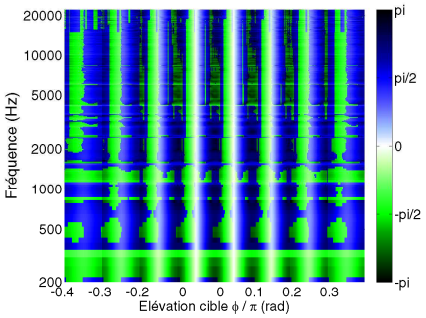
Erreur E_{75} (°) :



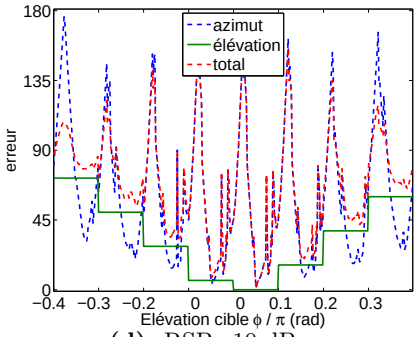
(a) RSB=0 dB.



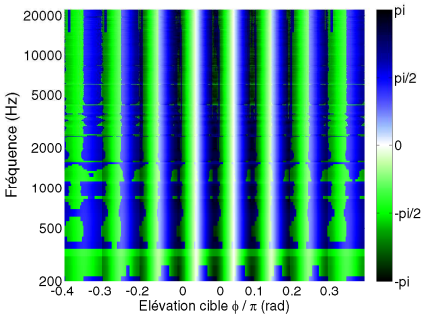
(b) RSB=0 dB.



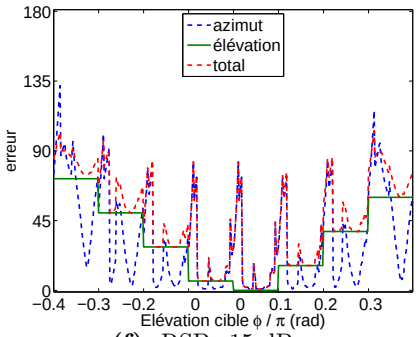
(c) RSB=10 dB.



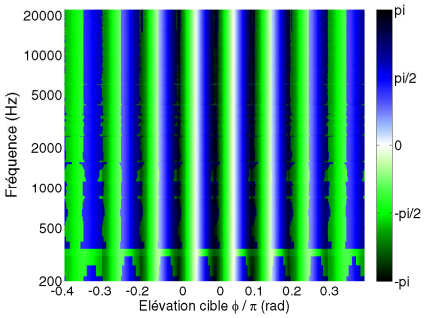
(d) RSB=10 dB.



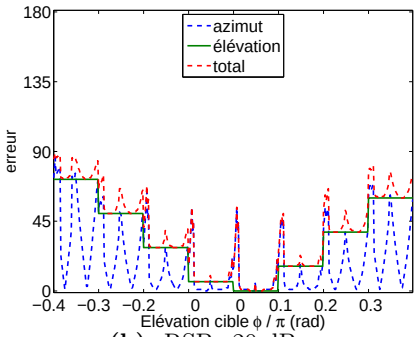
(e) RSB=15 dB.



(f) RSB=15 dB.



(g) RSB=20 dB.



(h) RSB=20 dB.

Figure IV.18 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source perturbatrice à la direction (0,0) à différents RSB (ligne 1 0 dB, ligne 2 10 dB, ligne 3 15 dB, ligne 4 20 dB). L'estimation de l'azimut (colonne 1) est affiché dans une représentation conjointe position-fréquence. L'erreur E_{75} associée est représentée dans la colonne de droite. Localisation effectuée avec un couple microphonique cardioïde composé de deux capsules cardioïdes écartées de 2 cm selon l'axe $\vec{y}$.

Localisation avec trois microphones Afin d'estimer la valeur de l'élévation, un troisième capteur cardioïde est inséré pointant vers $\vec{z}$, selon la configuration illustrée en figure IV.12. Les résultats de la localisation sont présentés dans la figure IV.19. Comme énoncé précédemment (IV.3.1.d), les erreurs occasionnées par les interférences destructives lors de l'addition des signaux pour l'obtention de S_0 engendrent une erreur de localisation. Cette erreur particulière intervient pour les fréquences supérieures à 5 kHz, et principalement pour les azimuts proches de $\pm\pi/2$, donc lorsque τ_{12} est maximal.

Ce phénomène impacte la localisation en élévation en premier lieu, car cette valeur fait intervenir $S_0 = S_1 + S_2$ dans l'équation (IV.34). Pour mémoire, l'élévation ϕ est estimée par la forme

$$\phi = \sin^{-1}\left(\frac{2S_3}{S_0} - 1\right). \tag{IV.47}$$

D'après les relations (IV.40) et (IV.43), S_0 est alors soumis à un filtrage en peigne, pour des fréquences dont la longueur d'onde λ vérifie

$$\lambda < c\,\tau_{12}. \tag{IV.48}$$

Dans le cas où le filtrage en peigne engendre des faibles valeurs $S_0 < S_3$, l'équation (IV.47) devient

$$\phi = \sin^{-1}(\varsigma) \quad \text{avec} \quad \varsigma > 1. \tag{IV.49}$$

Comme la fonction $\sin^{-1}(\varsigma)$ est définie pour $\varsigma \in [-1, 1]$, ϕ ne peut pas être estimé.

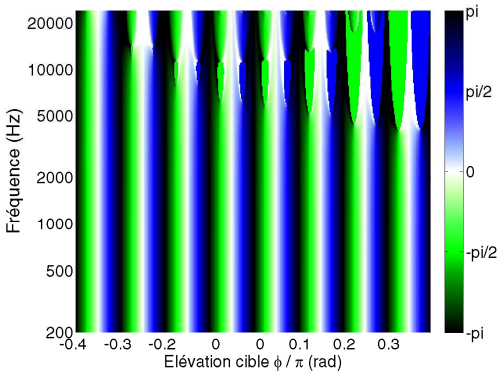
Ce phénomène apparaît ponctuellement pour des faibles élévations (autour de $\theta \approx \pm\pi/2$) et s'étale lorsque la source se décale vers le pôle nord. Effectivement, aux faibles élévations, la directivité M_3 induit des valeurs proches de 0 sur S_3. Ainsi, près du pôle sud, les faibles valeurs induites par le filtrage en peigne sur S_0 sont compensées par les faibles valeurs du signal S_3 ($S_0 > S_3$) amenant ς à l'intervalle $[-1, 1]$ où $\sin^{-1}$ est définie.

Étant donné que l'élévation et S_0 sont utilisés dans l'estimation de l'azimut (équation (IV.44)), la localisation sur cette coordonnée est également perturbée.

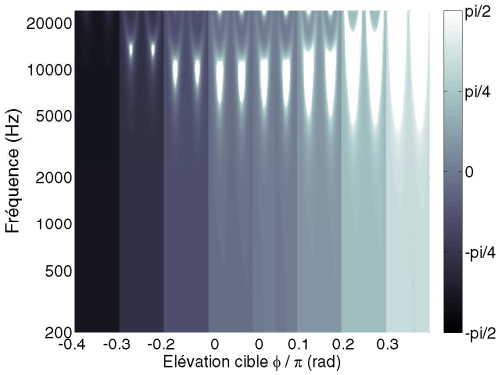
Ce problème peut être résolu en utilisant l'énergie des signaux pour constituer la référence S_0. Comme l'illustre la figure IV.20, la direction de la source est alors parfaitement estimée pour l'ensemble des fréquences, car le filtrage en peigne est ainsi évité.

En présence d'une source perturbatrice , l'élévation est correctement estimée à partir d'un RSB de 10 dB, avec un E_{75}, moyen pour cette coordonnée proche de 7° (figure IV.21). Pour ce même RSB, l'azimut n'est au contraire correctement évalué que lorsque la source est proche du plan horizontal, et pour des fréquences comprises entre 1 kHz et 10 kHz. L'indicateur E_{75} est alors dégradé et atteint une valeur proche de 45°. En effet, la présence d'une source perturbatrice dégrade la phase des signaux qui permet de résoudre l'ambigüité gauche-droite. Avec un RSB de 15 dB, la localisation est correcte. Seuls quelques sauts d'ambiguïté gauche-droite restent présents pour des fréquences inférieures à 500 Hz, permettant d'atteindre un E_{75} moyen total de 15° .

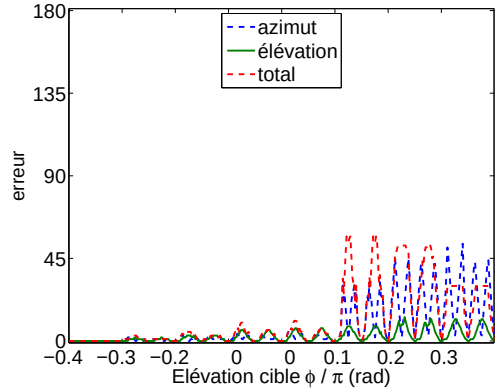
Sur la figure IV.22, on remarque que la position de la source perturbatrice en élévation n'a pas d'influence marquée sur la localisation. En effet, on observe uniquement une légère augmentation des dégradations lorsque la source perturbatrice s'écarte de l'axe $\vec{x}$, car elle perturbe l'information de phase qui permet la localisation hémisphérique. Pour ces positions de la source perturbatrice, l'indicateur E_{75} est de 17° contre 15° quand la source se trouve sur l'axe $\vec{x}$.



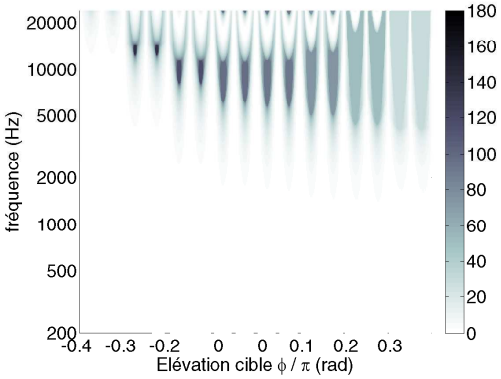
(a) Azimut θ.



(b) Elévation ϕ.

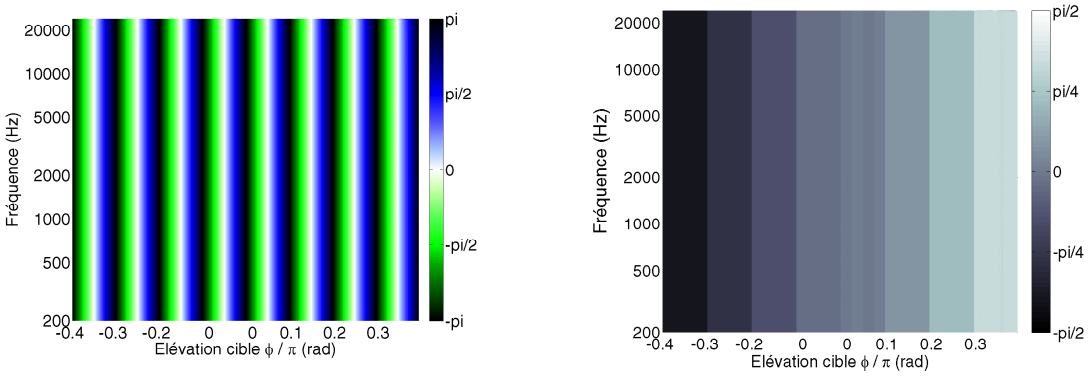


(c) Erreur $E75$ (°).



(d) Erreur E_t (°).

Figure IV.19 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14. Estimation de l'azimut(a), de l'élévation (b) et erreurs E_{75} (c) et E_t (d) associées. Résultats obtenus avec l'utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$ respectivement. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Le signal de pression considéré est $S_0 = S_1 + S_2$. Les figures (a), (b) et (d) sont affichées dans une représentation conjointe position-fréquence.

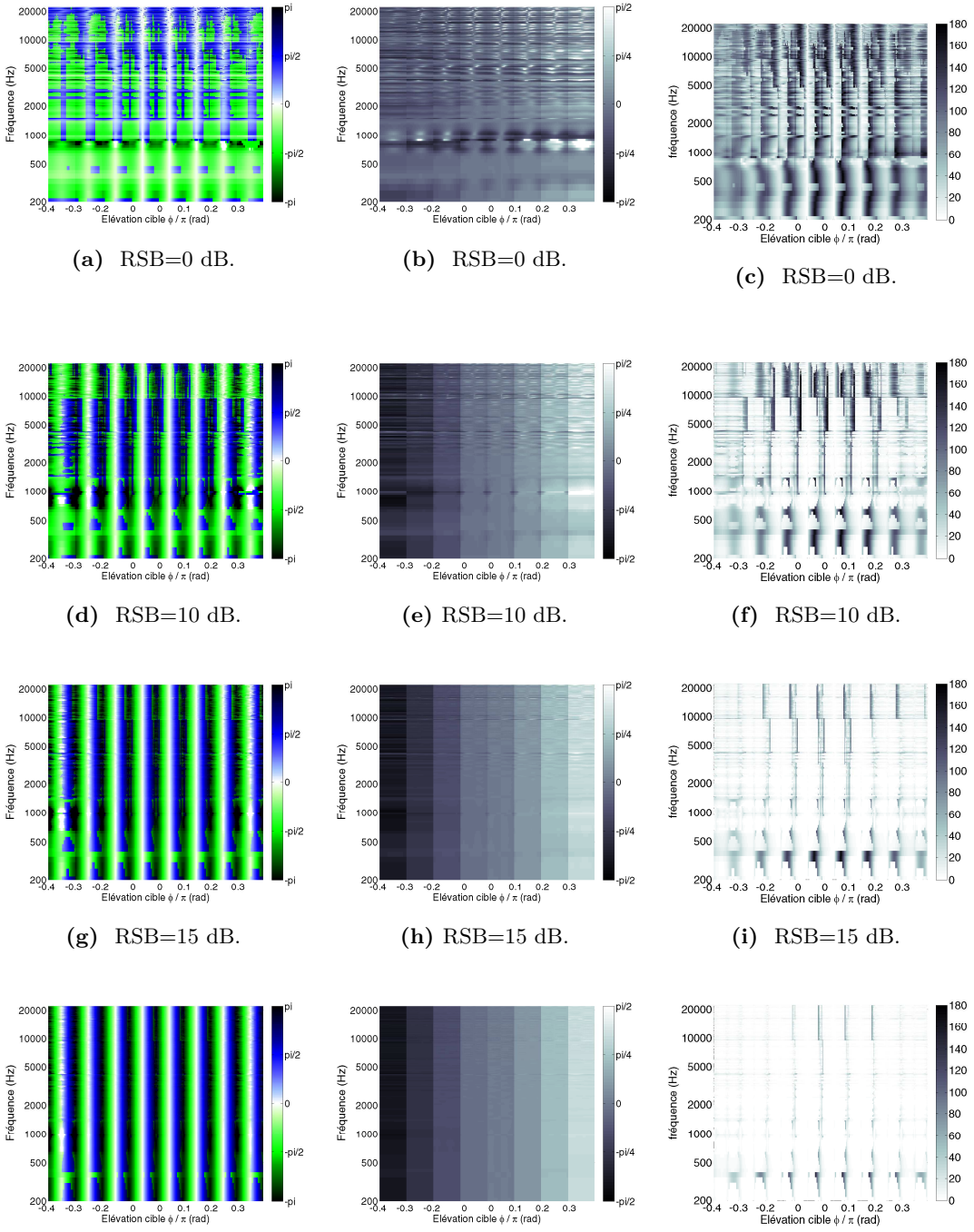


(a) Azimut θ.

(b) Elévation ϕ.

Figure IV.20 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14. L'estimation de l'azimut (a) et de l'élévation (b) est affichée dans une représentation conjointe position-fréquence. Résultats obtenus avec l'utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$ respectivement. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Le signal de pression considéré est $|S_0| = |S_1| + |S_2|$.

Azimut θ : Elévation ϕ : Erreur E_t (°) :



(a) RSB=0 dB. **(b)** RSB=0 dB. **(c)** RSB=0 dB.

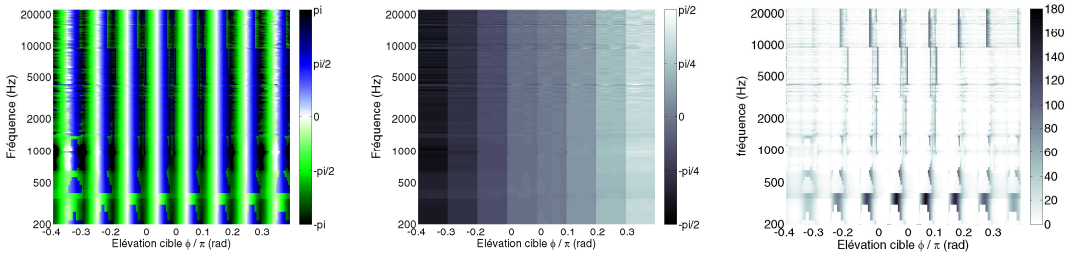
(d) RSB=10 dB. **(e)** RSB=10 dB. **(f)** RSB=10 dB.

(g) RSB=15 dB. **(h)** RSB=15 dB. **(i)** RSB=15 dB.

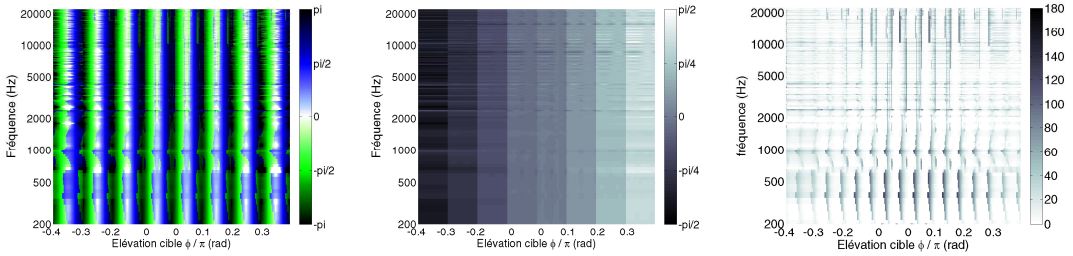
(j) RSB=20 dB. **(k)** RSB=20 dB. **(l)** RSB=20 dB.

Figure IV.21 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source perturbatrice à la direction (0,0) à différents RSB (ligne 1 0 dB, ligne 2 10 dB, ligne 3 15 dB, ligne 4 20 dB). L'estimation de azimut (colonne 1), de l'élévation (colonne 2) et de l'erreur E_t associée (colonne 3) sont affichées dans une représentation conjointe position-fréquence. Localisation effectuée avec l'utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Le signal de pression considéré est $|S_0| = |S_1| + |S_2|$.

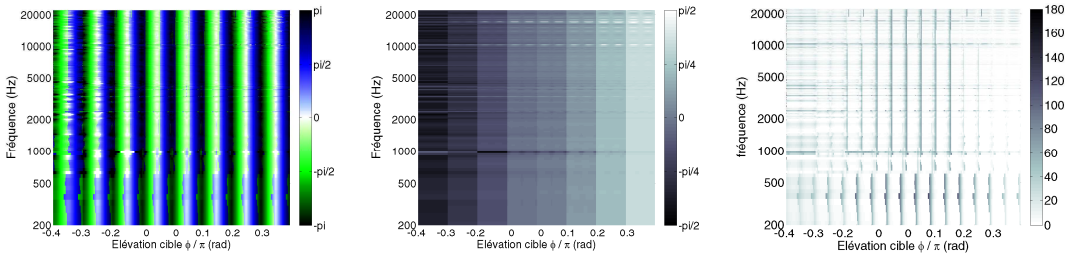
Azimut θ : Elévation ϕ : Erreur E_t (°) :



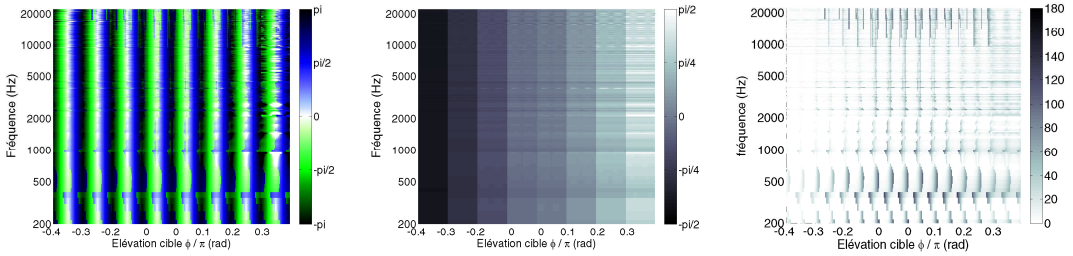
(a) Bruit à $(0, 0)$. **(b)** Bruit à $(0, 0)$. **(c)** Bruit à $(0, 0)$.



(d) Bruit à $(\pi/2, 0)$. **(e)** Bruit à $(\pi/2, 0)$. **(f)** Bruit à $(\pi/2, 0)$.



(g) Bruit à $(0, \pi/2)$. **(h)** Bruit à $(0, \pi/2)$. **(i)** Bruit à $(0, \pi/2)$.



(j) Bruit à $(0, -\pi/2)$. **(k)** Bruit à $(0, -\pi/2)$. **(l)** Bruit à $(0, -\pi/2)$.

Figure IV.22 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source perturbatrice à un RSB=15 dB à des positions différentes (ligne 1 (0,0), ligne 2 $(\pi/2, 0)$, ligne 3 $(0, \pi/2)$, ligne 4 $(0, -\pi/2)$). L'estimation de azimut (colonne 1), de l'élévation (colonne 2) et l'erreur E_t associée (colonne 3) sont affichées dans une représentation conjointe position-fréquence. Localisation effectuée avec l'utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Le signal de pression considéré est $|S_0| = |S_1| + |S_2|$.

IV.4 Première variante du prototype mettant en œuvre deux microphones bidirectionnels et un microphone cardioïde

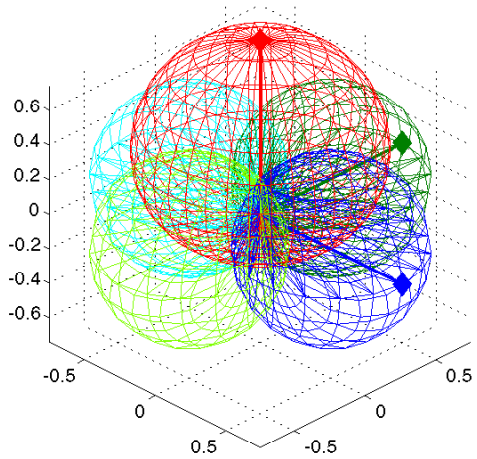


Figure IV.23 : Représentation des directivités microphoniques M_1, M_2 et M_3 dans l'espace 3D d'un capteur coïncidant composé de 2 capsules bidirectionnelles pointant respectivement vers $\vec{x}$ et $\vec{y}$ et d'une capsule cardioïde dirigée vers $\vec{z}$.

Nous avons voulu tester le principe de localisation utilisant d'autres configurations microphoniques, en cherchant à obtenir une directivité microphonique virtuelle s'approchant de la configuration définie.

IV.4.1 Synthèse de microphones cardioïdes virtuels

Comme nous l'avons évoqué en IV.3.1, le couple bidirectionnel présente une faible variation directionnelle dans sa configuration initiale. Nous cherchons à profiter des caractéristiques directionnelles d'un couple cardioïde, en les synthétisant à partir des signaux issus des microphones bidirectionnels. De la même manière que pour le système retenu, le troisième capteur cardioïde est aussi utilisé pour la localisation en élévation.

Les signaux $S_{1bi}, S_{2bi}, S_{3ca}$, issus de cette antenne microphonique, répondent à la relation

(IV.3) avec les paramètres de directivité suivants

$$\begin{cases} M_{1bi} & = \cos\theta\cos\phi \\ M_{2bi} & = \sin\theta\cos\phi \\ M_{3ca} & = \frac{1}{2}(1+\sin\phi) \end{cases}. \tag{IV.50}$$

Dans cette configuration, le signal S_o est obtenu par

$$S_o(t,\theta,\phi) = \frac{S_{1bi}^2(t,\theta,\phi) + S_{2bi}^2(t,\theta,\phi) + 4S_{3ca}^2(t,\theta,\phi)}{4S_{3ca}(t,\theta,\phi)}, \tag{IV.51}$$

et les signaux délivrés par les microphones cardioïdes virtuels S_{1caV} et S_{2caV} s'écrivent alors

$$\begin{cases} S_{1caV}(t,\theta,\phi) & = \frac{1}{2}(1+S_{1bi}(t,\theta,\phi)) \\ S_{2caV}(t,\theta,\phi) & = \frac{1}{2}(1-S_{1bi}(t,\theta,\phi)) \end{cases}. \tag{IV.52}$$

Les signaux issus des microphones cardioïdes virtuels peuvent être directement utilisés dans les équations (IV.31) et (IV.34), permettant ainsi de localiser la source en élévation et en azimut. Cette configuration étant coïncidante, une ambiguïté gauche-droite est à nouveau présente.

IV.4.2 Performances de localisation

Les performances de localisation sont étudiées selon le même protocole que celui défini en IV.3.3 et dans la figure IV.14.

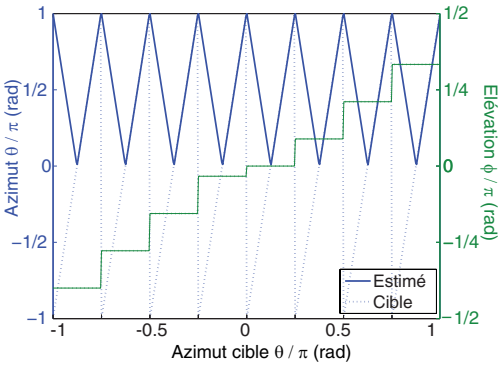
Les résultats de cette analyse sont illustrés par la figure IV.24. On y observe que la localisation en élévation est parfaite, mais l'azimut présente une ambiguïté gauche-droite due au fait de la coïncidence des capteurs.

Comme la localisation est effectuée en passant par des capteurs virtuels, effectuer un décalage physique des capteurs ne permettrait pas d'obtenir l'information manquante permettant de résoudre cette ambiguïté. L'utilisation de cette solution pourrait être envisagée lors de la captation d'une scène sonore se limitant à un demi-espace, comme par exemple une pièce de théâtre enregistrée depuis le public.

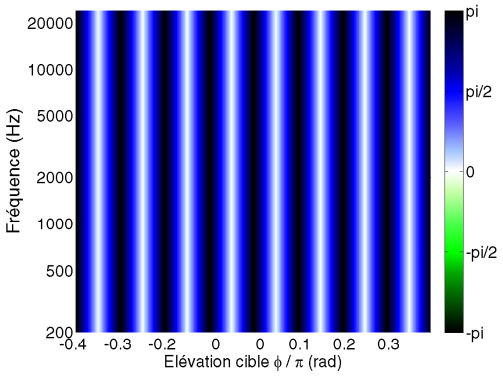
En présence d'une source perturbatrice localisée en (0,0) avec un RSB de 0 dB, les résultats de la localisation en azimut sont limités à $[0, \pi/2]$ (figure IV.25a). En effet, la source est repérée en un point intermédiaire entre sa position et celle de la source perturbatrice. D'autre part, cette configuration perturbe la localisation en élévation, plaçant la source de manière aléatoire lorsque celle-ci s'éloigne du plan horizontal (figure IV.25b). Lorsque le RSB s'approche de 10 dB, la source est bien localisée en azimut et en élévation, et l'indice E_{75} atteint des valeurs voisines de $20°$ dès que la source se trouve dans le bon hémisphère ($\theta \in [0, \pi]$), comme l'illustrent les figures IV.25d, IV.25e et IV.25f. En effet, l'ambiguïté gauche-droite persiste dégradant ainsi les performances. Lorsque le RSB est supérieur à 15 dB, seules quelques erreurs de localisation sont relevées pour les sources proches du pôle sud. Ces erreurs sont dues au fait qu'elles se trouvent au point "sourd" des trois microphones utilisés (figures IV.25).

La position de la source perturbatrice ne joue pas un rôle marqué lorsqu'elle est placée sur le plan horizontal. En revanche, lorsqu'elle se déplace en élévation, la localisation de cette coordonnée est dégradée, impactant directement l'estimation de la position globale de la source. En effet, comme l'illustre la figure IV.26, l'indicateur $E75$ atteint des valeurs comprises entre $5°$ et $45°$ et ces erreurs diminuent lorsque la source se déplace vers le pôle nord ($\phi \approx \pi/2$). Ce comportement est caractéristique et directement lié à la directivité cardioïde du microphone pointant vers l'axe $\vec{z}$, tel qu'énoncé en IV.3.1.c.

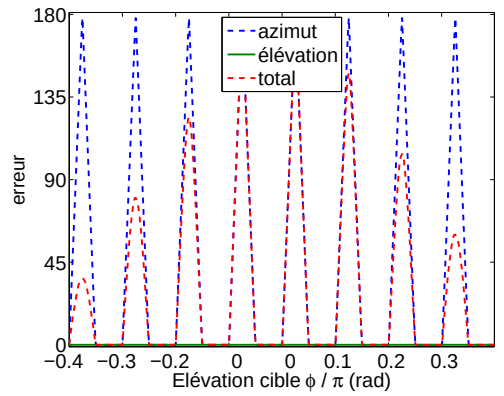
La solution étudiée ici est cependant difficile à mettre en œuvre. Les contraintes physiques empêchent d'une part de placer toutes les capsules microphoniques au même point. D'autre part, la conception de capteurs bidirectionnels n'est pas aisée car la membrane doit être libre sur ses deux faces, limitant ainsi la possibilité de méthodes de transduction à utiliser. Une recherche des capteurs bidirectionnels disponibles sur le marché nous a dissuadés de leur utilisation, car il existe très peu de modèles, qui plus est très onéreux. Cet aspect, est loin d'être négligeable lorsqu'il s'agit de la conception d'un outil grand public.



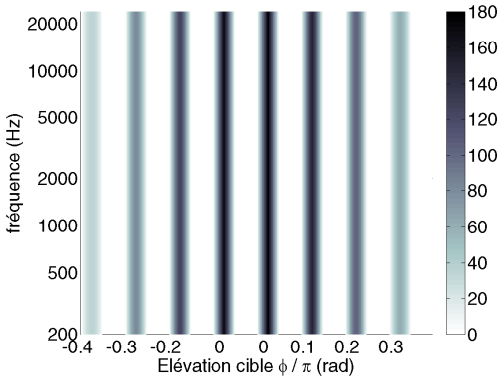
(a) Azimut θ (bleu), Élévation ϕ (vert) estimés à 1 kHz.



(b) Représentation conjointe position-fréquence de l'estimation de l'azimut θ.



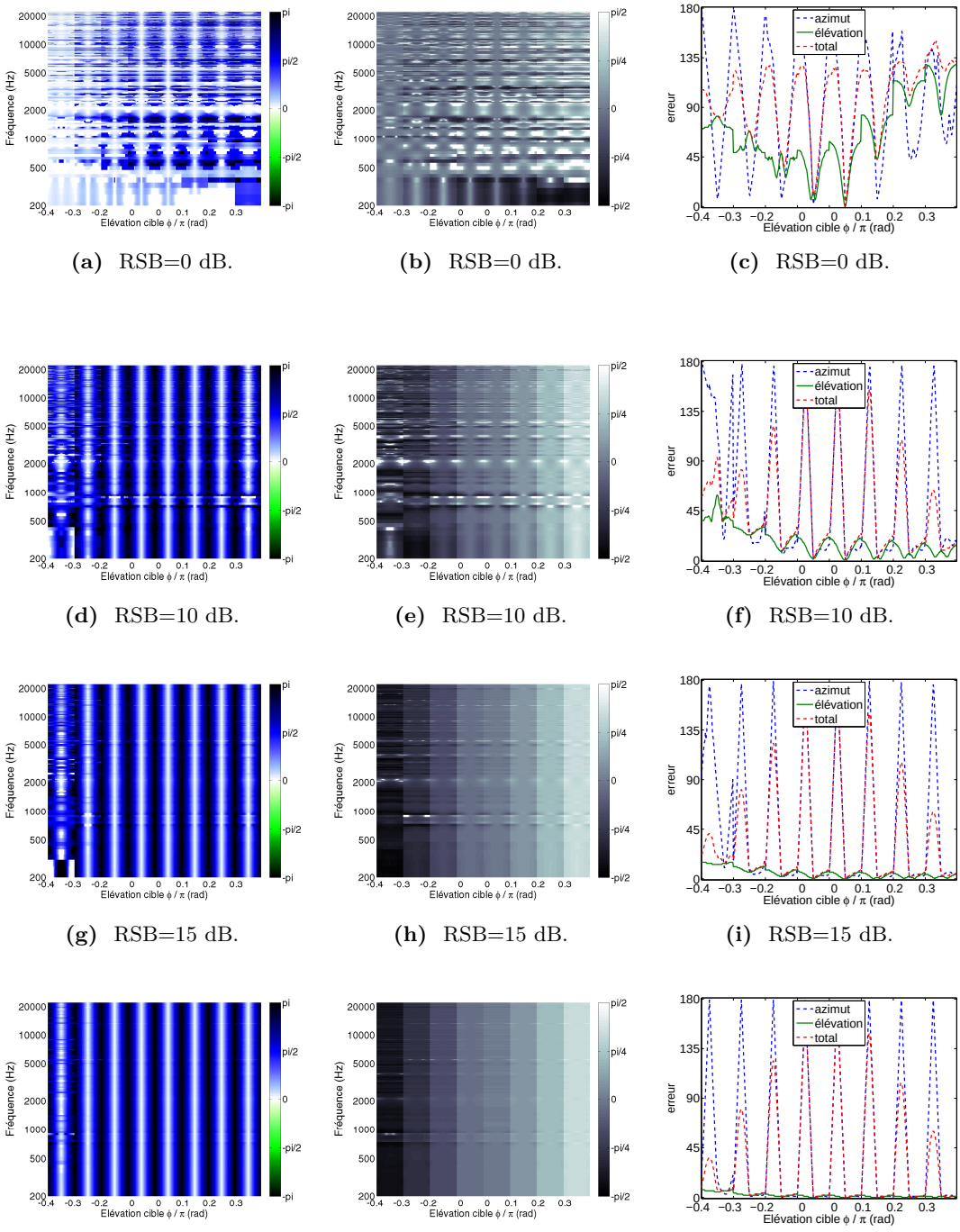
(c) Erreur E_{75} (°).



(d) représentation conjointe position-fréquence de l'erreur E_t

Figure IV.24 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 avec l'utilisation d'une antenne coïncidante composée de deux capsules bidirectionnelles et une cardioïde pointant respectivement vers $\vec{x}$, $\vec{y}$ et $\vec{z}$. (a) Estimation de la localisation en azimut (bleu) et élévation (vert) pour une fréquence $f = 1\ kHz$ en fonction de la position de la source cible. (b) Estimation de la localisation en élévation en fonction de la fréquence et la position cible de la source. (c) Erreur d'estimation E_{75} en fonction de la position cible de la source. (d) Erreur d'estimation E_t en fonction de la fréquence et la position cible de la source.

Azimut θ : Élévation ϕ : Erreur E_{75} (°) :



(a) RSB=0 dB. **(b)** RSB=0 dB. **(c)** RSB=0 dB.

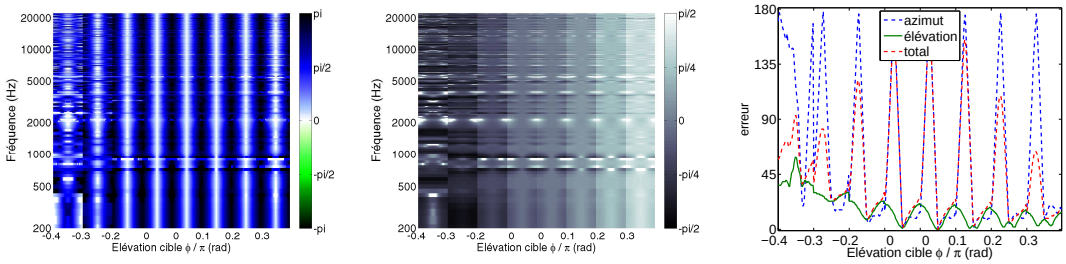
(d) RSB=10 dB. **(e)** RSB=10 dB. **(f)** RSB=10 dB.

(g) RSB=15 dB. **(h)** RSB=15 dB. **(i)** RSB=15 dB.

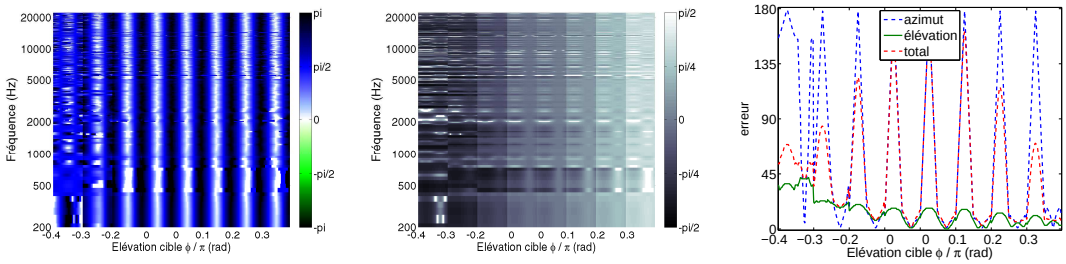
(j) RSB=20 dB. **(k)** RSB=20 dB. **(l)** RSB=20 dB.

Figure IV.25 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source de bruit à la direction (0,0) pour des différents RSB (ligne 1 0 dB, ligne 2 10 dB, ligne 3 15 dB, ligne 4 20 dB). Localisation effectuée avec l'utilisation d'une antenne microphonique coïncidante composée de 2 capteurs bidirectionnels pointant vers $\vec{x}$ et $\vec{y}$ respectivement et un capteur cardioïde pointant $\vec{z}$. L'estimation de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne.

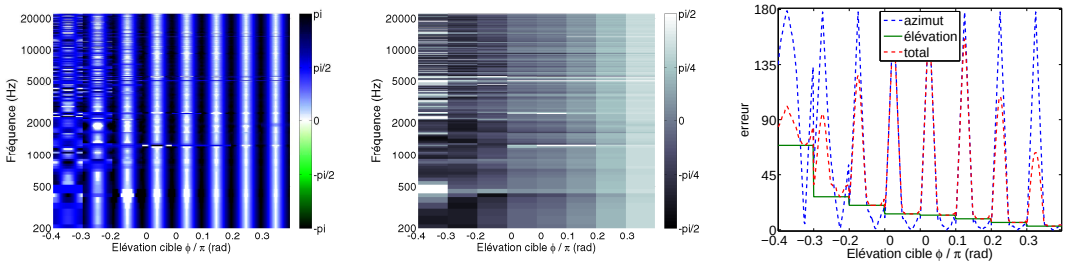
Azimut θ : Élévation ϕ : Erreur E_{75} (°) :



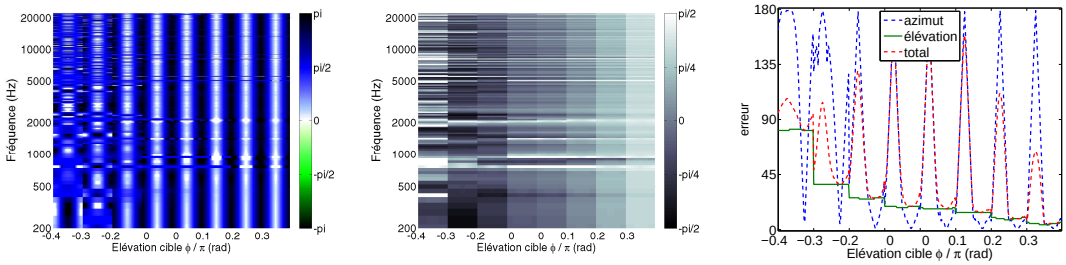
(a) Bruit à $(0, 0)$. **(b)** Bruit à $(0, 0)$. **(c)** Bruit à $(0, 0)$.



(d) Bruit à $(\pi/2, 0)$. **(e)** Bruit à $(\pi/2, 0)$. **(f)** Bruit à $(\pi/2, 0)$.



(g) Bruit à $(0, \pi/2)$. **(h)** Bruit à $(0, \pi/2)$. **(i)** Bruit à $(0, \pi/2)$.



(j) Bruit à $(0, -\pi/2)$. **(k)** Bruit à $(0, -\pi/2)$. **(l)** Bruit à $(0, -\pi/2)$.

Figure IV.26 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source de bruit à un RSB=15 dB à différentes positions (ligne 1 (0,0), ligne 2 $(\pi/2, 0)$, ligne 3 $(0, \pi/2)$, ligne 4 $(0, -\pi/2)$). Localisation effectuée avec l'utilisation d'une antenne microphonique coïncidante composée de 2 capteurs bidirectionnels pointant vers $\vec{x}$ et $\vec{y}$ respectivement et un capteur cardioïde pointant $\vec{z}$. L'estimation de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne.

IV.5 Seconde variante du prototype mettant en œuvre le format B de l'ambisonique à l'ordre 1

La méthode présentée en IV.4 n'est qu'un premier pas pour envisager l'utilisation de la méthode de localisation avec un microphone ambisonique à l'ordre 1.

IV.5.1 Synthèse des microphones cardioïdes virtuels

Les signaux X, Y et Z, issus du format B de l'ambisonique, peuvent être considérés comme trois microphones bidirectionnels orientés respectivement selon les axes $\vec{x}$, $\vec{y}$ et $\vec{z}$. Leur directivité est donc déterminée par la relation (IV.3), avec les paramètres de directivité suivants

$$\begin{cases} M_X &= \cos\theta\cos\phi \\ M_Y &= \sin\theta\cos\phi \\ M_Z &= \sin\phi \end{cases} . \tag{IV.53}$$

Le signal W, ou composante omnidirectionnelle, peut être directement considéré comme le signal de pression S_o.

Dans ce cas, les signaux composant le système microphonique virtuel comportant les trois capsules cardioïdes, pointant respectivement vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$, sont obtenus suivant les relations (IV.52),

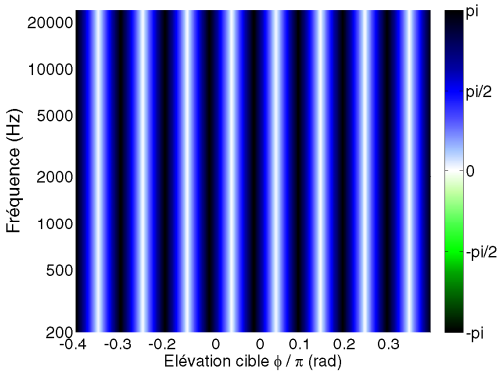
$$\begin{cases} S_{1caV}(t,\theta,\phi) &= \frac{1}{2}(1 + X(t,\theta,\phi)) \\ S_{2caV}(t,\theta,\phi) &= \frac{1}{2}(1 - X(t,\theta,\phi)) \\ S_{3caV}(t,\theta,\phi) &= \frac{1}{2}(1 - Z(t,\theta,\phi)) \end{cases} . \tag{IV.54}$$

Les signaux issus de ces microphones cardioïdes virtuels peuvent donc être utilisés directement dans les équations (IV.31) et (IV.34), permettant de localiser la source en élévation et en azimut. Cette configuration étant coïncidante, l'ambiguïté gauche-droite demeure.

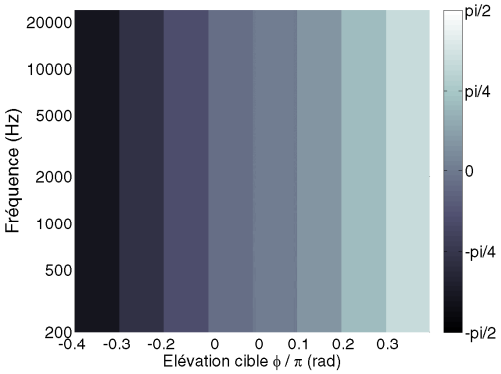
IV.5.2 Performances de localisation

Afin d'évaluer les performances de l'algorithme de localisation avec l'utilisation du format B de l'ambisonique, nous avons utilisé les paramètres définis en IV.3.3. Comme l'illustre la figure IV.27, les résultats sont similaires à ceux présentés en IV.4.2, obtenus à partir d'un couple bidirectif accompagné d'un microphone cardioïde (IV.4). Ainsi, on observe la persistance de l'ambiguïté gauche-droite ($\theta \in [0, \pi]$) due à la coïncidence des capsules.

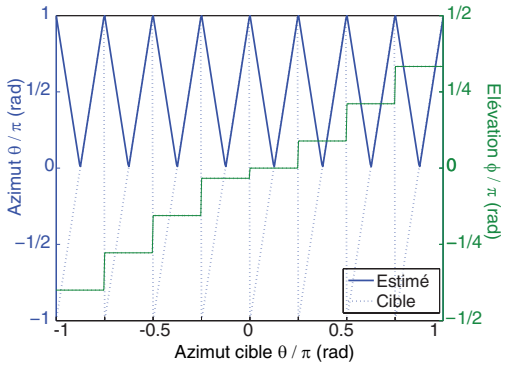
En présence d'une source perturbatrice, la localisation est dégradée et une source est localisée en un point intermédiaire, entre sa position et celle de la source perturbatrice. Lorsque le RSB est supérieur à 10 dB (figure IV.28), l'indice $E75$ est toujours inférieur à $30°$ quand la source se trouve dans le bon hémisphère ($\theta \in [0, \pi]$) et il décroît lorsque la source s'approche de la source perturbatrice (figure IV.29). Ce même phénomène est observé quelle que soit la position de la source perturbatrice à cause de la symétrie du capteur selon tous les axes.



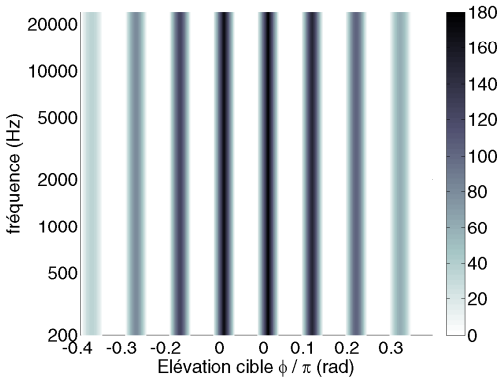
(a) Représentation conjointe position-fréquence de l'éstimation de azimut θ.



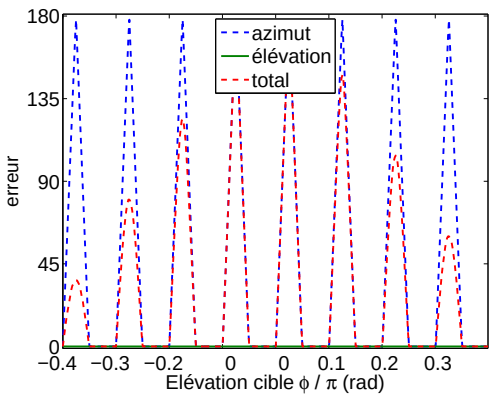
(b) Représentation conjointe position-fréquence de l'éstimation de l'élévation ϕ.



(c) Azimut θ (bleu), Élévation ϕ (vert) estimés à 1 kHz.



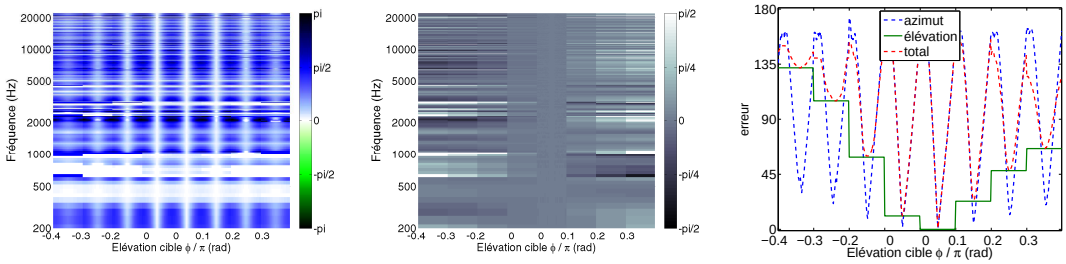
(d) Représentation conjointe position-fréquence de l'erreur E_t (°).



(e) Erreur E_{75} (°).

Figure IV.27 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 avec du format B de l'ambisonique. Localisation en azimut (a) et élévation (b) dans une représentation conjointe position-fréquence. (c) Localisation en azimut (bleu) et élévation (vert) pour une fréquence $f = 1\ kHz$ (c) en fonction de la position cible de la source. (d) Erreur d'estimation E_t en fonction de la fréquence et la position cible de la source. (e) Erreur d'estimation E_{75} en fonction de la position cible de la source.

Azimut θ : Élévation ϕ : Erreur E_{75} (°) :



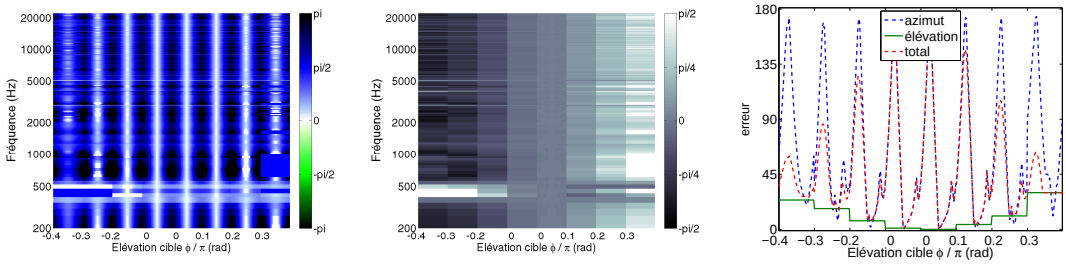
(a) RSB=0 dB. **(b)** RSB=0 dB. **(c)** RSB=0 dB.

(d) RSB=10 dB. **(e)** RSB=10 dB. **(f)** RSB=10 dB.

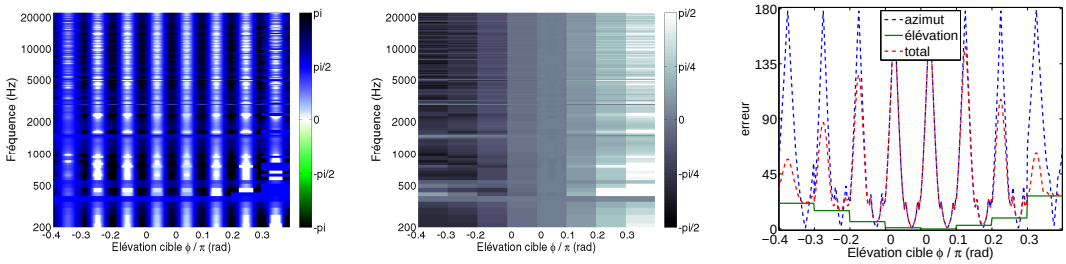
(g) RSB=15 dB. **(h)** RSB=15 dB. **(i)** RSB=15 dB.

Figure IV.28 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source de bruit à (0,0) avec des RSB différents (ligne 1 0 dB, ligne 2 10 dB, ligne 3 15 dB). Localisation effectuée à partir du format B de l'ambisonique à l'ordre 1. L'estimation de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne.

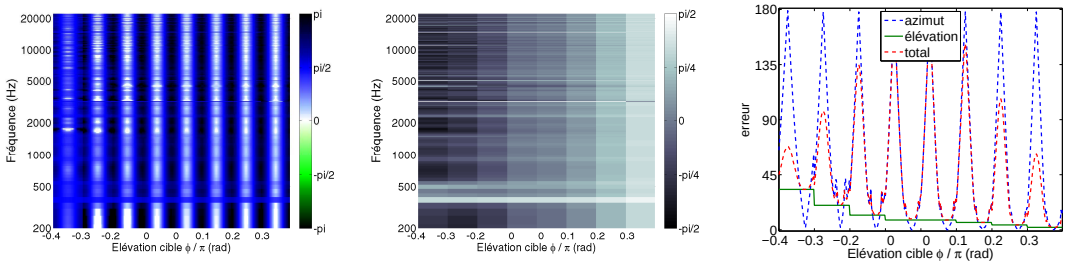
Azimut θ : Élévation ϕ : Erreur E_{75} (°) :



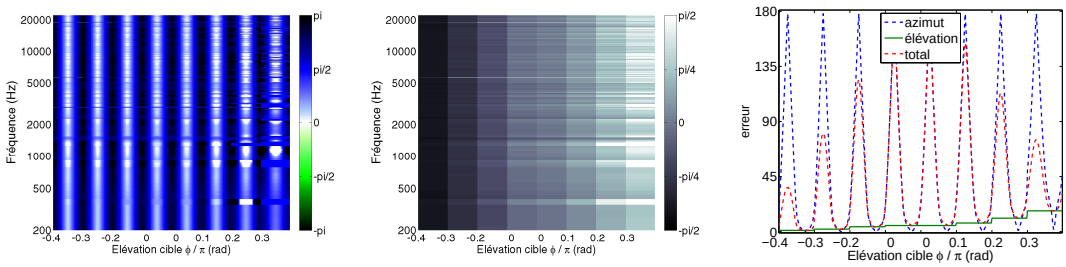
(a) θ, Bruit à $(0, 0)$. **(b)** Bruit à $(0, 0)$. **(c)** Bruit à $(0, 0)$.



(d) Bruit à $(\pi/2, 0)$. **(e)** Bruit à $(\pi/2, 0)$. **(f)** Bruit à $(\pi/2, 0)$.



(g) Bruit à $(0, \pi/2)$. **(h)** Bruit à $(0, \pi/2)$. **(i)** Bruit à $(0, \pi/2)$.



(j) Bruit à $(0, -\pi/2)$. **(k)** Bruit à $(0, -\pi/2)$. **(l)** Bruit à $(0, -\pi/2)$.

Figure IV.29 : Localisation d'une source large bande suivant la trajectoire illustrée dans la figure IV.14 en présence d'une source de bruit avec un RSB=10 dB à des directions différentes (ligne 1 (0,0), ligne 2 $(\pi/2, 0)$, ligne 3 $(0, \pi/2)$, ligne 4 $(0, -\pi/2)$). Localisation effectuée à partir du format B de l'ambisonique à l'ordre 1. L'estimation de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne.

IV.6 Conclusion

Cette étude a démontré la faisabilité de la mise en œuvre d'un système de captation utilisant uniquement trois capteurs pour la localisation d'une source dans l'espace 3D. Grâce à l'utilisation de la directivité des capteurs, un indice directionnel est disponible pour extraire l'information spatiale de la scène sonore.

Tout d'abord, l'étude des caractéristiques géométriques des différentes directivités a permis de définir la meilleure configuration microphonique avec une analyse du gradient directionnel des capteurs. L'utilisation d'une antenne microphonique composée de trois microphones cardioïdes a été définie comme la meilleure configuration microphonique pour permettre une bonne localisation des sources. Cette configuration présente une ambiguïté hémisphérique qui a été résolue avec le décalage de deux microphones placés sur le plan horizontal.

Le post-traitement étant effectué dans le domaine fréquentiel, une position est obtenue pour chaque fréquence et à chaque trame temporelle. Cette analyse permet également l'utilisation d'une moyenne énergétique des signaux, et non leur amplitude, éliminant ainsi les interférences destructives induites par le décalage des capteurs.

Deux configurations microphoniques complémentaires ont été étudiées, confirmant ainsi la possibilité de l'utilisation de la méthode à partir du format B de l'ambisonique à l'ordre 1 et d'un dispositif composé de deux microphones bidirectifs. Ces deux dispositifs microphoniques étant coïncidents, ils ne permettent pas la discrimination hémisphérique, et une erreur de localisation persiste lorsque l'azimut de la source se trouve entre 0 et $-\pi$.

Toute cette étude est basée sur l'hypothèse de capteurs idéaux, avec des directivités contrôlées associées à des réponses fréquentielles plates. Même si ces hypothèses peuvent être vérifiées sur une bande de fréquences limitée et en absence de tout objet diffractant (dispositif mobile ou main de l'utilisateur), il convient d'adapter cette méthode de localisation aux caractéristiques réelles des microphones intégrés dans leur support.


Phone
Mail
Safari
iPod

V Caractéristiques microphoniques réelles dans la localisation des sources : concept d'Ob-RTF

Dans le chapitre précédent, nous avons défini une configuration microphonique ainsi qu'un algorithme permettant la localisation de sources. Ce dernier, en partant de l'hypothèse que les microphones présentent des caractéristiques idéales en termes de directivité et de réponse en fréquence, n'est valable que sur la plage de fréquences où les microphones respectent ces caractéristiques. De plus, lorsque les capteurs sont introduits dans un dispositif diffractant, les caractéristiques microphoniques sont à nouveau modifiées, ce qui limite d'autant plus les performances de localisation.

Afin de pallier ce problème, nous souhaitons prendre en compte ces paramètres comme des indices supplémentaires permettant la localisation des sources.

A ce titre, nous souhaitons généraliser l'algorithme présenté au chapitre précédent, en effectuant une description détaillée des réponses directives dans tout l'espace des capteurs utilisés *in situ*. Cette représentation, très proche des HRTF, nous conduit à étudier le lien entre ces deux paramètres, introduisant une généralisation du concept des HRTF, l'*Object Related Transfert Function* ou fonction de transfert liée à l'objet (Ob-RTF), quel que soit l'objet.

V.1 Localisation sonore basée sur les HRTF

Les HRTF, ou fonctions de transfert liées à la tête, sont les réponses en fréquence directionnelles décrivant le trajet acoustique entre une source sonore et l'entrée du conduit auditif, en fonction de la direction de la provenance de l'onde. Dans le domaine temporel, elles sont décrites par leurs réponses impulsionnelles ou HRIR (I.2). Partant de ce principe, MacDonald [MacDonald, 2008] et Keyrous [Keyrouz and Diepold, 2006, Keyrouz et al., 2006] considèrent que les HRTF contiennent l'ensemble de l'information spatiale permettant le décodage spatial d'une scène auditive.

Keyrous et Diepold [Keyrouz and Diepold, 2006] ont d'abord proposé d'appliquer une déconvolution sur les signaux $s_l(t)$ et $s_r(t)$ captés respectivement par les oreilles gauche et droite, avec l'ensemble des HRIR mesurées sur une sphère. Les signaux produits par une source placée à (θ_s, ϕ_s) et captés par les oreilles correspondent en effet à

$$s_l(t) = s_0(t) * HRIR_l(t, \theta_s, \phi_s) \tag{V.1a}$$

et

$$s_r(t) = s_0(t) * HRIR_r(t, \theta_s, \phi_s), \tag{V.1b}$$

où $*$ représente le produit de convolution et $s_0(t)$ correspond au signal de pression acoustique pouvant être mesuré à la position du centre de la tête (sujet absent).

Il est donc possible d'obtenir le signal $s_0(t)$ en appliquant le filtre inverse noté $HRIR^{-1}(t\theta_s, \phi_s)$. Comme la direction de la source est inconnue, il est nécessaire d'effectuer la déconvolution des signaux s_l et s_r avec l'ensemble des directions (θ, ϕ) des HRIR de l'oreille opposée. Cette relation est exprimée par

$$\begin{aligned} \gamma_l(t, \theta, \phi) &= s_l(t) * HRIR_l^{-1}(t, \theta, \phi) \\ &= s_0(t) * HRIR_l(t, \theta_s, \phi_s) * HRIR_l^{-1}(t, \theta, \phi) \end{aligned} \tag{V.2a}$$

et

$$\begin{aligned} \gamma_r(t, \theta, \phi) &= s_r(t) * HRIR_r^{-1}(t, \theta, \phi) \\ &= s_0(t) * HRIR_r(t, \theta_s, \phi_s) * HRIR_r^{-1}(t, \theta, \phi), \end{aligned} \tag{V.2b}$$

la direction (t, θ_s, ϕ_s) pour laquelle les deux quantités $\gamma_l(t, \theta_s, \phi_s)$ et $\gamma_r(t, \theta_s, \phi_s)$ sont égales correspondant à celle de la source,

$$s_l(t) * HRIR_l^{-1}(t, \theta, \phi) = s_r(t) * HRIR_r^{-1}(t, \theta, \phi) = s_0(t) \quad \text{(V.3)}$$

ssi $\theta = \theta_s$ et $\phi = \phi_s$.

Une fonction d'intercorrélation $\star$ est ensuite appliquée entre les deux sous-ensembles des signaux résultants

$$\gamma_l(t, \theta, \phi) \star \gamma_r(t, \theta, \phi), \quad \text{(V.4)}$$

permettant de retrouver l'égalité (V.3) et par la suite d'estimer la position de la source.

Ce procédé permet ainsi de localiser une source unique et immobile sur l'intervalle temporel étudié.

Plus récemment, MacDonald [MacDonald, 2008] a proposé une adaptation de la méthode de Keyrous dans le domaine fréquentiel.

Le signal de pression à l'oreille gauche est caractérisé par

$$S_L(f, \hat{\theta}, \hat{\phi}) = S_0(f)\ HRTF_L(f, \hat{\theta}, \hat{\phi}), \quad \text{(V.5)}$$

où $S_0(f)$ est la réponse en fréquence du signal de pression acoustique $s_0(t)$ défini précédemment, et où $HRTF_L(f, \hat{\theta}, \hat{\phi})$ est la réponse en fréquence de l'oreille gauche pour une source placée à $(\hat{\theta}, \hat{\phi})$. Par la suite, les indices L et R désignent respectivement les grandeurs correspondant aux oreilles gauche et droite.

En multipliant le signal d'une oreille par la HRTF de l'oreille opposée, mesurée à la direction de la source, nous obtenons

$$\lambda_{LR}(f, \hat{\theta}, \hat{\phi}) = S_L(f, \hat{\theta}, \hat{\phi})\ HRTF_R(f, \hat{\theta}, \hat{\phi}) = S_o(f)\ HRTF_L(f, \hat{\theta}_s, \hat{\phi}_s)\ HRTF_R(f, \hat{\theta}, \hat{\phi}) \quad \text{(V.6a)}$$

et

$$\lambda_{RL}(f,\hat{\theta},\hat{\phi}) = S_R(f,\hat{\theta},\hat{\phi})\ HRTF_L(f,\hat{\theta},\hat{\phi}) = S_o(f)\ HRTF_R(f,\hat{\theta_s},\hat{\phi_s})\ HRTF_L(f,\hat{\theta},\hat{\phi}), \quad \text{(V.6b)}$$

d'où

$$\lambda_{LR}(f,\hat{\theta},\hat{\phi}) = \lambda_{RL}(f,\hat{\theta},\hat{\phi}) \quad \text{(V.7)}$$

ssi $\hat{\theta} = \hat{\theta_s}$ et $\hat{\phi} = \hat{\phi_s}$.

Pour identifier la direction de la source, il est donc nécessaire :

- d'effectuer le produit du signal avec l'ensemble des directions (θ, ϕ) pour lesquelles les HRTF sont connues,
- d'identifier la direction pour laquelle l'égalité (V.7) est vérifiée parmi cet ensemble.

Cette méthode permet d'estimer une localisation avec des erreurs inférieures à 5 ° en utilisant de 2 microphones, et des erreurs inférieures à 2 ° en utilisant 4 microphones pour un rapport signal à bruit proche de 40 dB [Keyrouz et al., 2006].

V.2 Analyse de scène sonore basée sur les Ob-RTF

Tout microphone présente une directivité ainsi qu'une réponse en fréquence intrinsèque à ses dimensions et à sa conception. Le champ sonore capté par celui-ci est par ailleurs modifié par les diffractions et réflexions engendrées par les objets qui l'entourent, tels que les préamplificateurs ou leur support. Lorsque le microphone est inséré ou intégré à un objet, sa réponse en fréquence et sa directivité sont modifiées, mais elles peuvent être mesurées. Pour généraliser le concept d'HRTF, nous définissons les *Object Related Transfert Function* ou fonction de transfert liée à l'objet (Ob-RTF), comme les réponses en fréquence d'un microphone dépendant de la fréquence et de la direction de la source en azimut θ et en élévation ϕ. Les Ob-RTF peuvent être mesurées de manière similaire aux HRTF en utilisant les méthodes existantes [Langendijk and Bronkhorst, 2000, Majdak et al., 2007, Pernaux, 2003]. Il s'agit principalement de mesurer la réponse du dispositif pour un ensemble de directions situées sur la sphère 3D. Dans le cas d'un microphone directionnel parfait, sa réponse est identique autour de son axe de révolution. Le signal émis par une source sonore est alors identique quelle que soit sa position sur le cercle défini autour de l'axe de révolution du microphone. Connaissant la directivité du microphone, ainsi que le signal capté par un microphone omnidirectionnel au même point (signal de référence), il est alors possible de déterminer la position du cercle contenant la source. Si un deuxième microphone est utilisé, la solution se limite à l'intersection des deux cercles. Une fois que le dispositif est inséré dans un objet, sa directivité et sa réponse en fréquence sont modifiées, on parle alors d'Ob-RTF. En fonction de la géométrie de l'objet, la symétrie de la directivité des microphones est conservée ou non. Dans le cas où la géométrie est cassée, connaissant les Ob-RTF ainsi que le signal de référence, l'analyse sur le principe de localisation par les HRTF décrit dans V.1 permet de déterminer une solution unique. Ainsi, l'algorithme de localisation proposé par MacDonald [MacDonald, 2008] et Keyrous [Keyrouz and Diepold, 2006, Keyrouz et al., 2006] offre la possibilité d'utiliser un jeu de HRTF afin d'obtenir une bonne localisation sans aucune connaissance *a priori* de la source à localiser. Nous proposons ici d'exploiter ces méthodes sans avoir recours à une captation binaurale, en s'affranchissant ainsi de la tête acoustique grâce aux Ob-RTF.

V.3 Les Ob-RTF dans la localisation des sources

Comme dans le chapitre IV, la méthode part de l'hypothèse qu'il n'existe qu'une seule source à chaque instant par bande fréquentielle. Le traitement s'effectue alors sur des fenêtres temporelles dont la taille doit être déterminée en fonction de l'écart des capteurs et en fonction du nombre d'échantillons fréquentiels souhaité. Il est également possible de rajouter des zéros (*zeropadding*), en fonction de la discrétisation spectrale souhaitée.

La relation (V.7) peut être généralisée à toute paire microphonique (m, n) en remplaçant les HRTF par les Ob-RTF, suivant

$$\begin{cases} \lambda_{LR}(f,\hat{\theta},\hat{\phi}) = S_m(f,\hat{\theta},\hat{\phi})\ ObRTF_n(f,\theta',\phi') \\ \lambda_{RL}(f,\hat{\theta},\hat{\phi}) = S_n(f,\hat{\theta},\hat{\phi})\ ObRTF_m(f,\theta',\phi') \end{cases}, \tag{V.8}$$

où $S_x(f,\hat{\theta},\hat{\phi})$ est le signal issu du capteur x et où $ObRTF_x(f,\theta',\phi')$ est la réponse en fréquence du capteur x pour la direction (θ',ϕ'). Comme précédemment, la relation (V.7) est vérifiée ssi

$$(\hat{\theta},\hat{\phi}) = (\theta',\phi').$$

Ce cas théorique n'est cependant pas toujours atteint. On cherche alors à trouver la meilleure estimation de la direction (θ',ϕ') parmi l'ensemble des directions (θ,ϕ). Celle-ci est obtenue en minimisant la distance δ entre les quantités λ_{LR} et λ_{RL}.

Considérons $\beta_{\hat{n},m}(f,\theta,\phi)$ comme le produit du signal $S_n(f,\hat{\theta},\hat{\phi})$, généré par une source placée à $(\hat{\theta},\hat{\phi})$, avec l'ensemble des $ObRTF_m(f,\theta,\phi)$ du microphone m, pour toutes les directions (θ',ϕ') disponibles suivant

$$\beta_{\hat{n},m}(f,\theta,\phi) = S_n(f,\hat{\theta},\hat{\phi})\ ObRTF_m(f,\theta,\phi). \tag{V.9}$$

$\beta_{\hat{n},m}$ est représenté par un vecteur de longueur L correspondant au nombre des directions

des Ob-RTF, de la forme,

$$\beta_{\hat{n},m}(f,\theta,\phi) = \begin{pmatrix} \beta_{\hat{n},m}(f,\theta_1,\phi_1) \\ \vdots \\ \beta_{\hat{n},m}(f,\theta_l,\phi_l) \\ \vdots \\ \beta_{\hat{n},m}(f,\theta_L,\phi_L) \end{pmatrix} = S_n(f,\hat{\theta},\hat{\phi}) \begin{pmatrix} ObRTF_m(f,\theta_1,\phi_1) \\ \vdots \\ ObRTF_m(f,\theta_l,\phi_l) \\ \vdots \\ ObRTF_m(f,\theta_L,\phi_L) \end{pmatrix}. \quad \text{(V.10)}$$

En écrivant $\beta_{\hat{n},m}(l) = \beta_{\hat{n},m}(f,\theta_l,\phi_l)$ et $\beta_{\hat{n},m}(l) = \beta_{\hat{n},m}(f,\theta_l,\phi_l)$, on définit la distance $\delta_{n,m}(l)$ entre chaque paire d'éléments $\beta_{\hat{n},m}(l)$ et $\beta_{\hat{m},n}(l)$

$$\delta_{n,m}(l) = \delta\left(\beta_{\hat{n},m}(l), \beta_{\hat{m},n}(l)\right). \quad \text{(V.11)}$$

On considère alors que la position de la source $(\hat{\theta},\hat{\phi})$ correspond à la direction où $\delta_{n,m}(l)$ est minimale.

Lorsque plus de deux capteurs sont utilisés simultanément, l'analyse s'effectue sur l'ensemble des paires microphoniques. La direction retenue est celle pour laquelle la valeur $\delta_{m,n}(l)$ est la plus faible sur l'ensemble de paires considérées.

V.4 Choix d'un critère de distance δ

Les données manipulées étant des valeurs complexes, la norme des vecteurs n'est pas le meilleur indicateur de ressemblance de ces derniers. Trois indicateurs de distance sont comparés par la suite.

Nous cherchons à définir la distance entre les quantités $A, B \in \mathbb{C}$ définies par

$$\begin{cases} A \equiv \beta_{\hat{n},m}(l) = S_m(f,\hat{\theta},\hat{\phi}) ObRTF_n(f,\theta_l,\phi_l) \\ B \equiv \beta_{\hat{n},m}(l) = S_n(f,\hat{\theta},\hat{\phi}) ObRTF_m(f,\theta_l,\phi_l) \end{cases} .$$

Distance angulaire : dans un premier temps, nous proposons d'utiliser le cosinus de l'angle entre les deux valeurs complexes comme indicateur de distance δ entre ces deux valeurs. Celui-ci est donc calculé grâce au produit scalaire,

$$D_{Ang.}(A,B) = \cos(\widehat{AB}) = \frac{A \cdot B}{|A||B|}. \tag{V.12}$$

Cette relation présente l'avantage de fournir un indice normalisé de ressemblance, qui vaut 0 lorsque les deux vecteurs sont colinéaires et qui vaut 1 lorsqu'ils sont perpendiculaires. Néanmoins, cette méthode ne rend pas compte de la norme des vecteurs.

Indice de Tanimoto : afin d'utiliser l'amplitude et l'angle des vecteurs dans un seul indicateur, un deuxième indicateur de distance $D_T(A,B)$ est proposé [Camacho and al, 2011], issu de l'indice de Tanimoto $T(A,B)$, défini par

$$D_T(A,B) = 1 - T(A,B) \tag{V.13}$$

$$= 1 - \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B} \tag{V.14}$$

où $\cdot$ définit le produit scalaire et $|\ |$ est la norme du vecteur.

Quotient : une dernière mesure de distance appelée méthode du quotient est proposée. Partant de l'équation (V.7), avec les paramètres de (V.8),

$$\frac{A}{B} = 1, \tag{V.15}$$

la distance $D_Q = \delta\left(\beta_{\hat{n},m}, \beta_{\hat{m},n}\right)$ est alors définie par

$$D_Q = \frac{A}{B} - 1. \tag{V.16}$$

L'avantage de cette distance est de ne calculer qu'une seule fois les rapports $\frac{ObRTF_n(f,\theta,\phi)}{ObRTF_m(f,\theta,\phi)}$ car ils sont constants tout au long de l'analyse, et sont indépendants des signaux $S_n(f,\theta,\phi)$. Cette distance permet de réduire considérablement le nombre d'opérations en allégeant ainsi la complexité.

V.5 Performances de localisation

Afin d'évaluer les performances de la méthode des Ob-RTF pour la localisation de sources, un protocole identique à celui décrit au chapitre IV a été utilisé. Le signal correspond à celui défini en IV.3.3 suivant la trajectoire illustrée en figure IV.14. Pour mémoire, il s'agit donc d'une source large bande tournant autour du microphone à des élévations différentes, par paliers de 20° partant de l'hémisphère sud vers le zénith. Pour permettre une évaluation comparative avec la méthode présentée dans le chapitre précédent, la configuration microphonique définie en IV.3.2 est retenue. Il s'agit donc de trois microphones pointant vers $\vec{x}$,$-\vec{x}$ et $-\vec{z}$, dont les deux premiers sont écartés de 2 cm sur l'axe $\vec{y}$.

Les analyses effectuées ici permettent de déterminer la meilleure méthode capable d'évaluer la distance δ (V.4) et la robustesse de l'algorithme de localisation en présence d'une source perturbatrice.

Les indicateurs utilisés ici sont ceux définis en IV.3.3.a.

V.5.1 Évaluation des indicateurs de distance

Les critères de distance définis en V.4 sont comparés. L'utilisation des critères de Tanimoto et du quotient permet une localisation correcte sans ambiguïté en l'absence de source perturbatrice (figures V.2 et V.3). L'utilisation de la distance angulaire laisse apparaître des nombreuses erreurs de localisation (figure V.1) où l'erreur moyenne totale E_{75} est de

52° (42° pour l'azimut et 32° pour l'élévation). Les erreurs sont plus marquées lorsque la source s'approche du zénith pour atteindre des erreurs maximales de localisation E_{75} d'environ 130°.

Comme les résultats obtenus à partir de l'indice de Tanimoto D_T ou du critère du quotient D_Q sont libres de toute erreur dans des conditions prises en compte ici (figures V.2 et V.3), nous limitons les études ultérieures à ces deux mesures de distance.

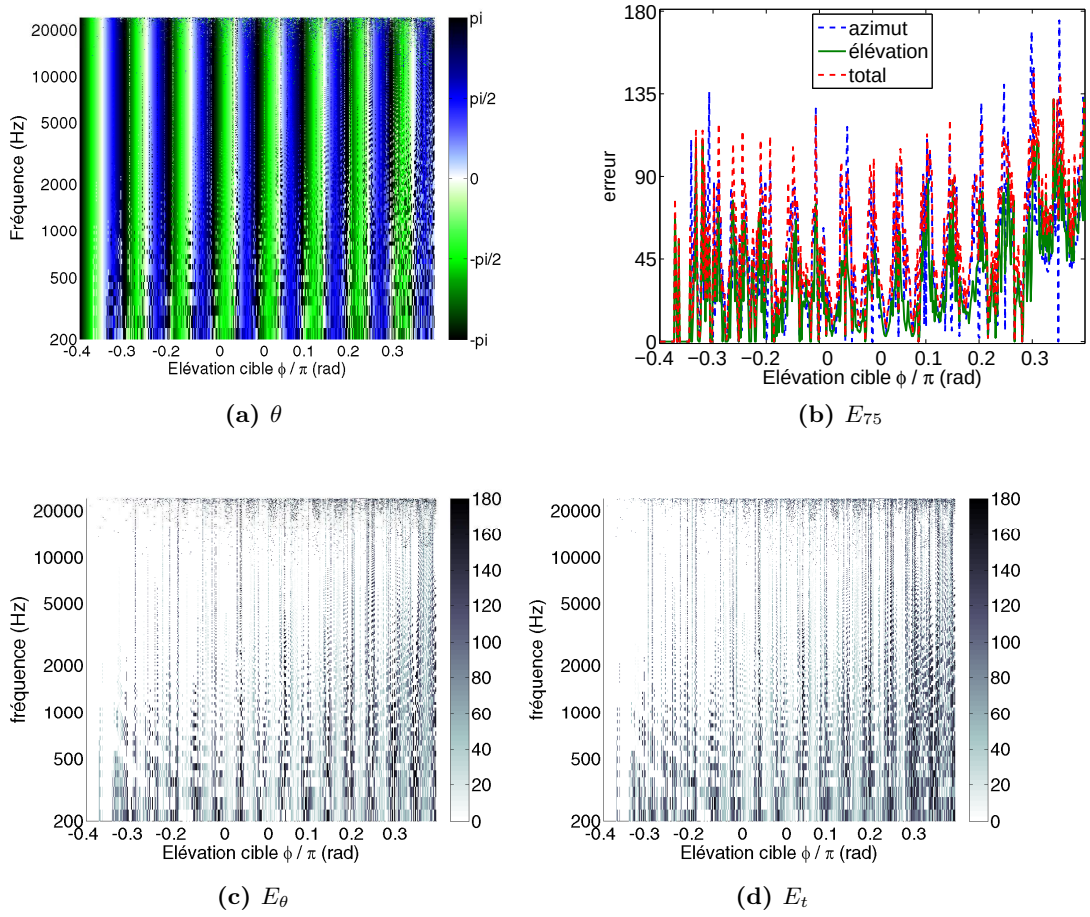
En présence d'une source perturbatrice, tel qu'illustré en figures V.4 et V.5, les critères de Tanimoto et du quotient donnent des résultats de localisation identiques quel que soit le rapport signal à bruit. Lorsque le RSB est supérieur à 30 dB, la source perturbatrice dégrade principalement l'information de phase. Comme cette donnée est critique pour la localisation des basses fréquences, les erreurs s'y focalisent, dégradant ainsi la qualité du E_{75} moyen. L'erreur E_{75} totale moyennée sur l'ensemble des directions est de 25° lorsque le RSB est de 20 dB, et atteint 3° pour un RSB de 30 dB. Cette erreur s'approche de 0° quand la source sonore se trouve au même azimut que la source perturbatrice, et est maximale lorsque ces deux sources sont diamétralement opposées.

La figure V.6 montre qu'il n'y a pas d'influence marquée de la position de la source perturbatrice sur les performances de localisation.

On observe également que les critères de Tanimoto et du quotient donnent des résultats équivalents pour la localisation des sources. Il est à noter que l'utilisation de la méthode du quotient permet de calculer la matrice $\frac{ObRTF_n(f,\theta,\phi)}{ObRTF_m(f,\theta,\phi)}$ (V.4) une seule fois et de la stocker en mémoire, ce qui réduit considérablement la complexité algorithmique. En effet, les calculs réalisés avec cette méthode ne nécessitent qu'entre 25 et 30% du temps utilisé par la méthode de Tanimoto.

V.5.2 Variante à deux capteurs

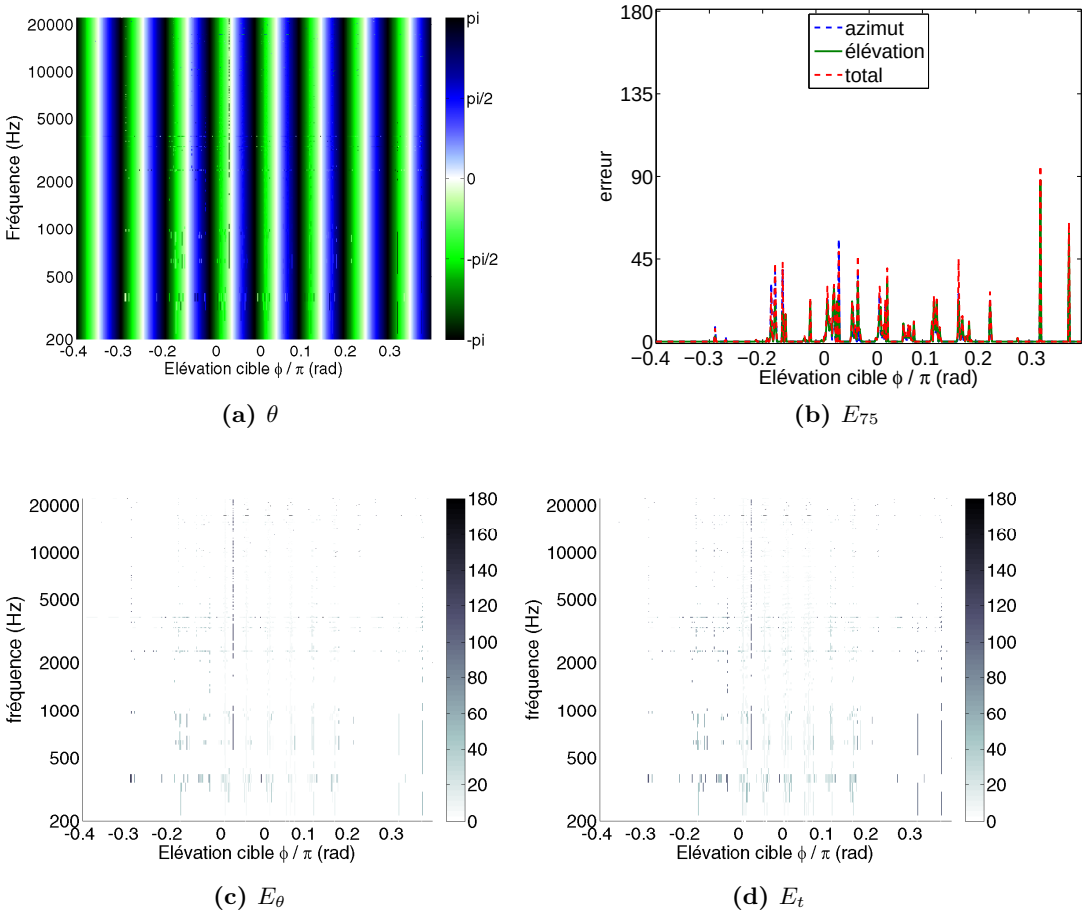
Nous considérons ici le cas où seuls les deux microphones placés sur le plan horizontal sont utilisés. En l'absence de toute source perturbatrice, il est possible de localiser les sources sur l'ensemble de l'espace 3D comme l'illustre la figure V.7. En présence d'une source perturbatrice, l'indicateur E_{75} moyen atteint 54°pour un RSB de 15 dB, 32° pour un RSB de 20 dB et 3° pour un RSB de 20 dB. Lorsque le RSB est de 40 dB, le E_{75} moyen sur l'azimut est de 0,3°, confortant ainsi les résultats de Keyruz et McDonald [Keyrouz et al., 2006, MacDonald, 2008].



(a) θ **(b)** E_{75}

(c) E_θ **(d)** E_t

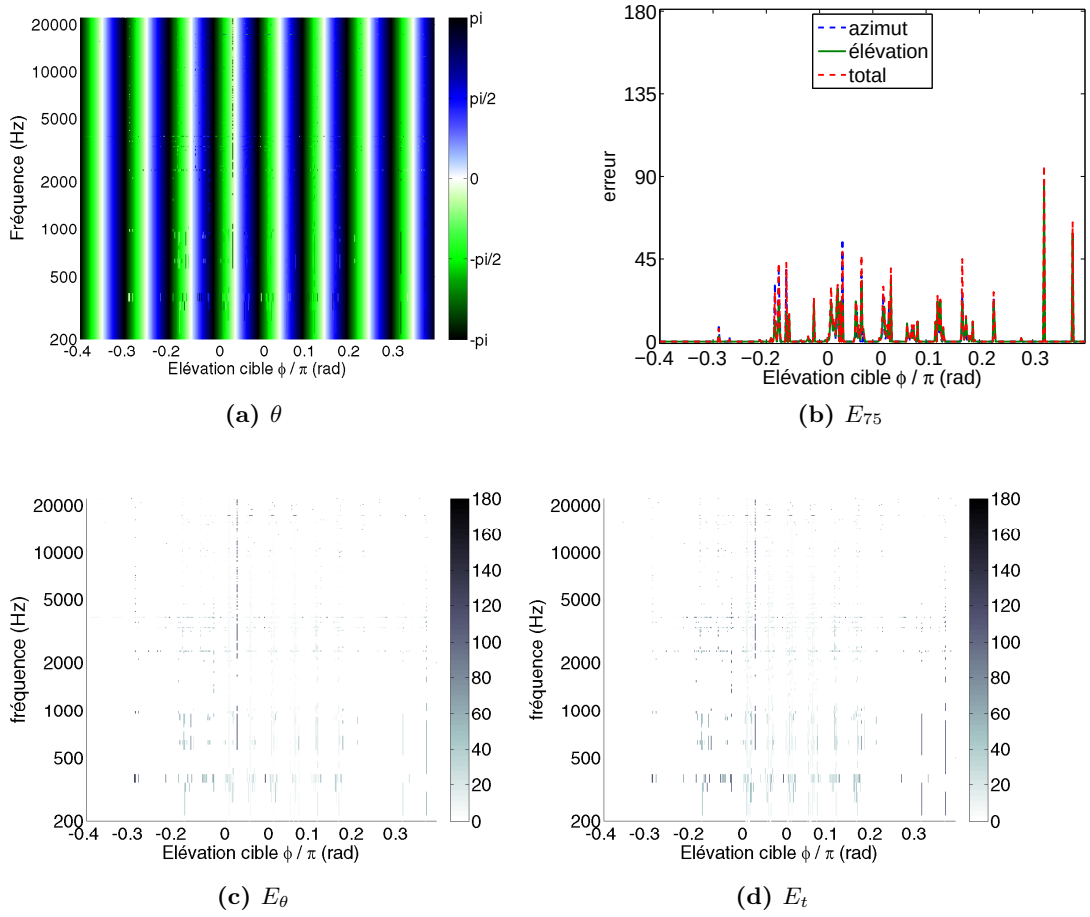
Figure V.1 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance $D_{ang.}$, d'une source large bande suivant la trajectoire illustrée en figure IV.14. Utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. (a) Estimation de l'azimut en fonction de la fréquence et la position cible de la source. (b) Erreur d'estimation E_{75} en fonction de la position cible de la source. Erreurs d'estimation E_θ (c) et E_t (d) associées affichées dans une représentation conjointe position-fréquence.



(a) θ **(b)** E_{75}

(c) E_θ **(d)** E_t

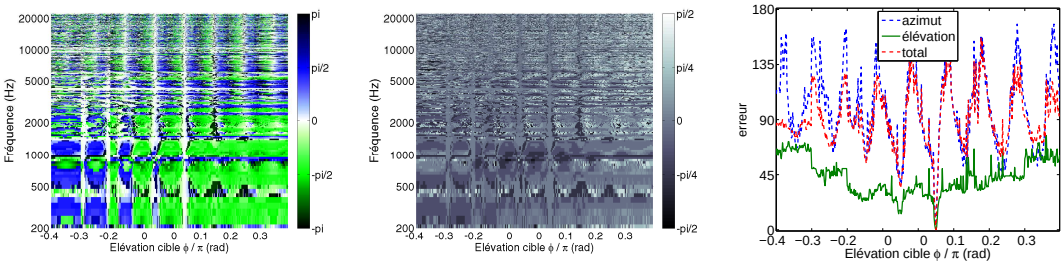
Figure V.2 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_T, d'une source large bande suivant la trajectoire illustrée en figure IV.14. Utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. (a) Estimation de l'azimut en fonction de la fréquence et la position cible de la source. (b) Erreur d'estimation E_{75} en fonction de la position cible de la source. Erreurs d'estimation E_θ (c) et E_t (d) associées affichées dans une représentation conjointe position-fréquence.



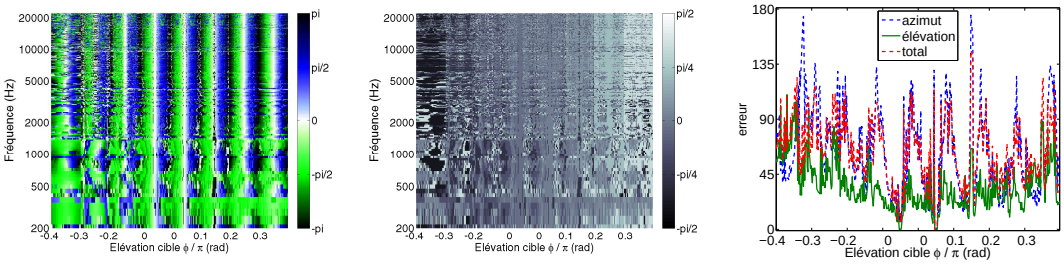
(a) θ **(b)** E_{75}

(c) E_θ **(d)** E_t

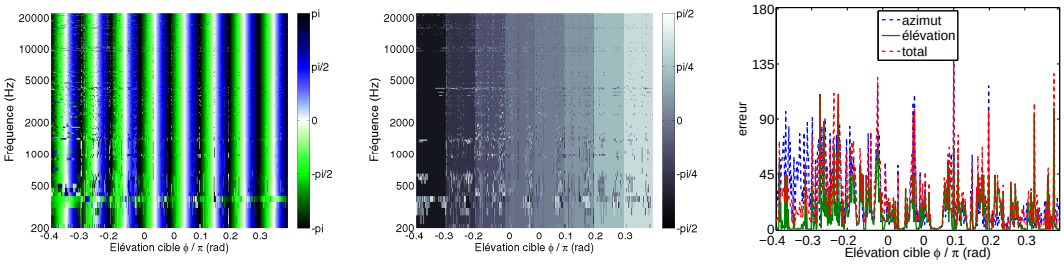
Figure V.3 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_Q, d'une source large bande suivant la trajectoire illustrée en figure IV.14. Utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. (a) Estimation de l'azimut en fonction de la fréquence et la position cible de la source. (b) Erreur d'estimation E_{75} en fonction de la position cible de la source. Erreurs d'estimation E_θ (c) et E_t (d) associées affichées dans une représentation conjointe position-fréquence.



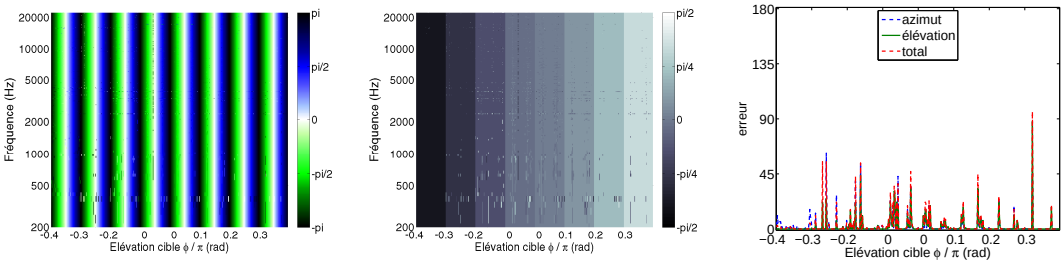
(a) θ : RSB=0 dB **(b)** ϕ : RSB=0 dB **(c)** E_{75} : RSB=0 dB



(d) θ : RSB=10 dB **(e)** ϕ : RSB=10 dB **(f)** E_{75} : RSB=10 dB

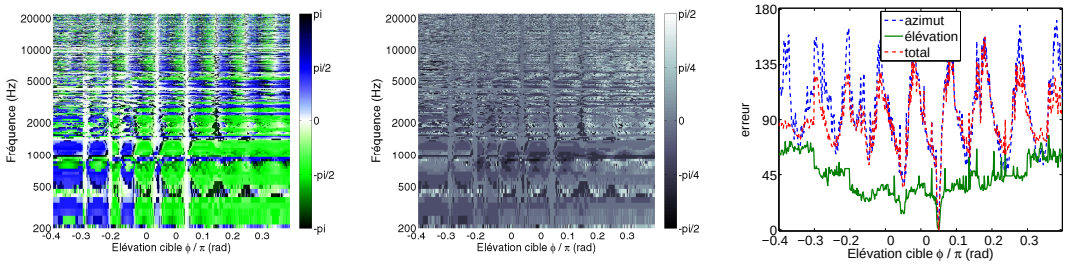


(g) θ : RSB=20 dB **(h)** ϕ : RSB=20 dB **(i)** E_{75} : RSB=20 dB

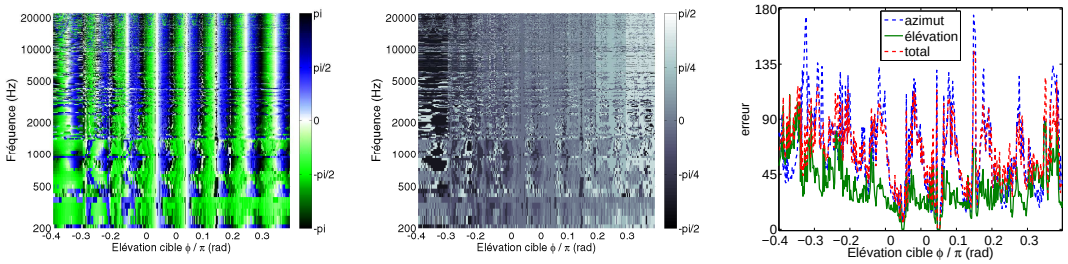


(j) θ : RSB=30 dB **(k)** ϕ : RSB=30 dB **(l)** E_{75} : RSB=30 dB

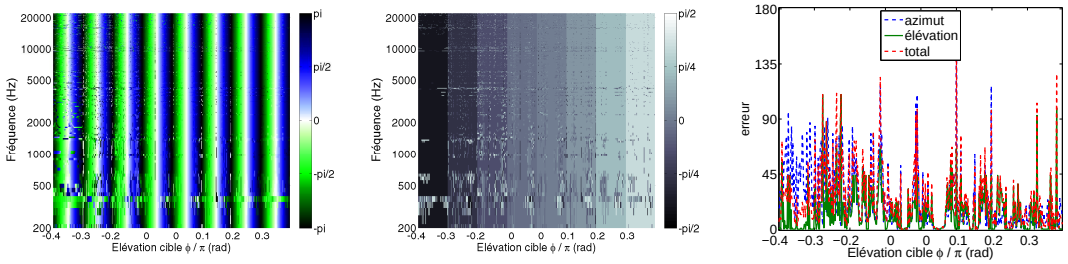
Figure V.4 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_Q, d'une source large bande suivant la trajectoire illustrée en figure IV.14 en présence d'une source de bruit à (0,0) et pour différents RSB (ligne 1 0 dB, ligne 2 10 dB, ligne 3 20 dB, ligne 4 30 dB). Utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Les estimations de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne.



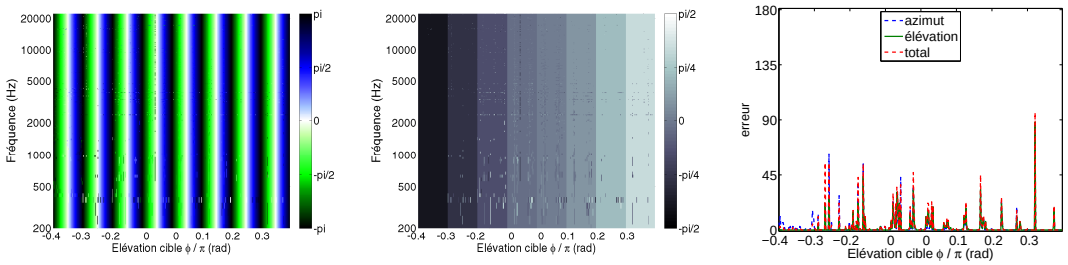
(a) θ : RSB=0 dB **(b)** ϕ : RSB=0 dB **(c)** E_{75} : RSB=0 dB



(d) θ : RSB=10 dB **(e)** ϕ : RSB=10 dB **(f)** E_{75} : RSB=10 dB

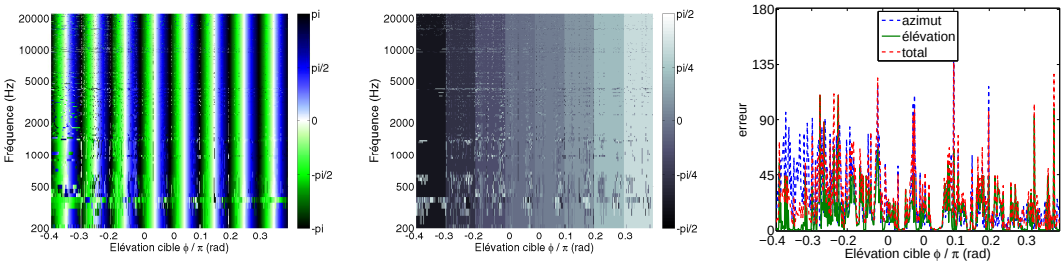


(g) θ : RSB=20 dB **(h)** ϕ : RSB=20 dB **(i)** E_{75} : RSB=20 dB

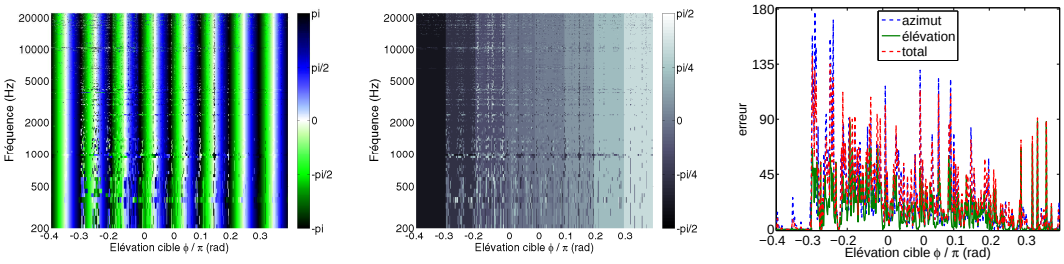


(j) θ : RSB=30 dB **(k)** ϕ : RSB=30 dB **(l)** E_{75} : RSB=30 dB

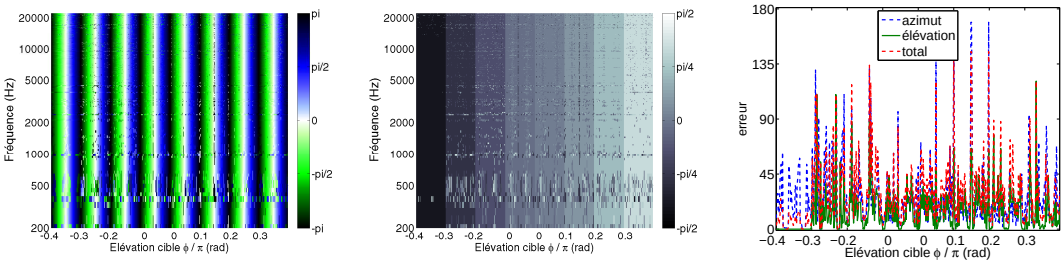
Figure V.5 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_T, d'une source large bande suivant la trajectoire illustrée en figure IV.14 en présence d'une source de bruit à (0,0) et pour différents RSB (ligne 1 0 dB, ligne 2 10 dB, ligne 3 20 dB, ligne 4 30 dB). Utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Les estimations de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne.



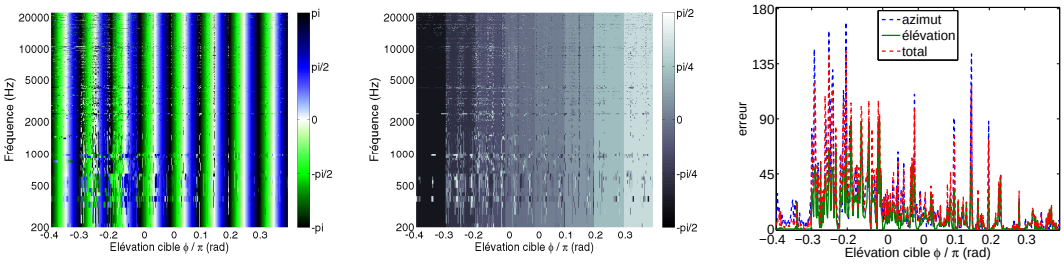
(a) θ : Bruit à (0,0) **(b)** ϕ : Bruit à (0,0) **(c)** E_{75} : Bruit à (0,0)



(d) θ : Bruit à (0,π/2) **(e)** ϕ : Bruit à (0,π/2) **(f)** E_{75} : Bruit à (0,π/2)

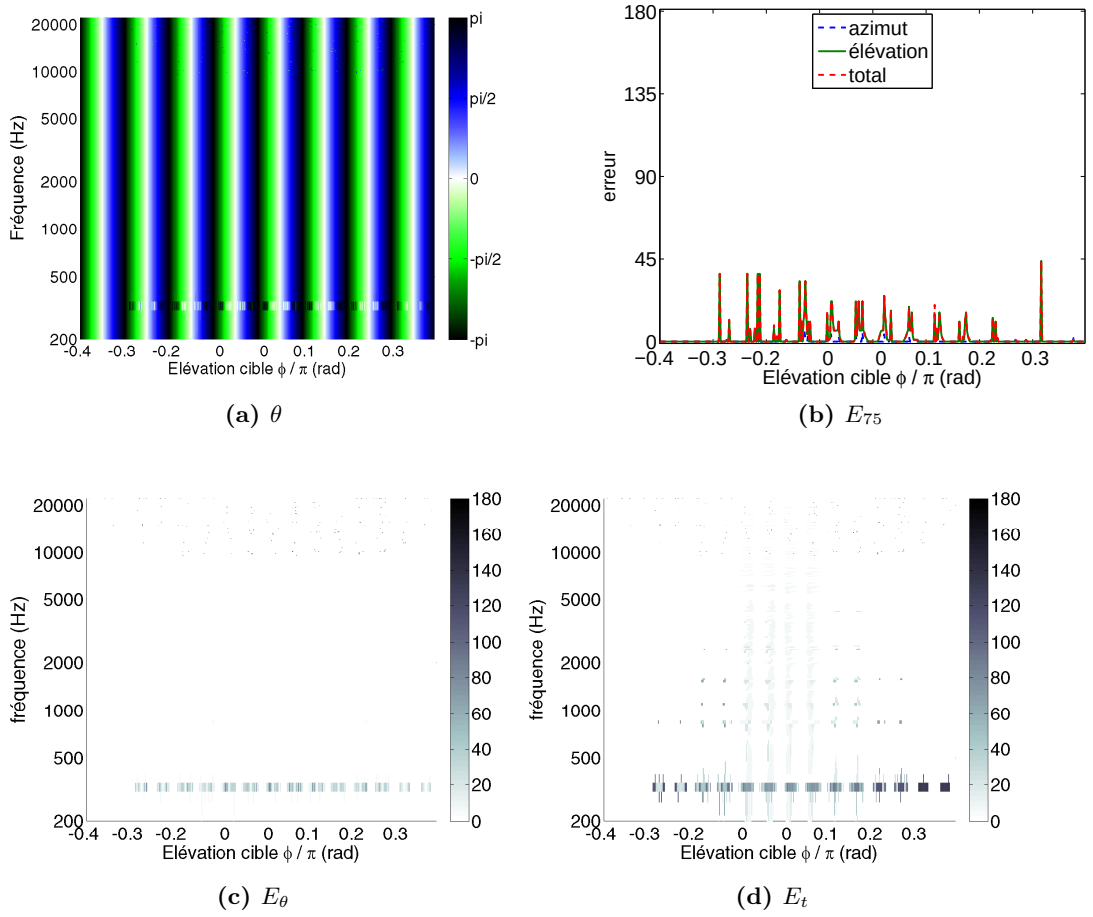


(g) θ : Bruit à (0,$-\pi$/2) **(h)** ϕ : Bruit à (0,$-\pi$/2) **(i)** E_{75} : Bruit à (0,$-\pi$/2)



(j) θ : Bruit à (π/2,0) **(k)** ϕ : Bruit à (π/2,0) **(l)** E_{75} : Bruit à (π/2,0)

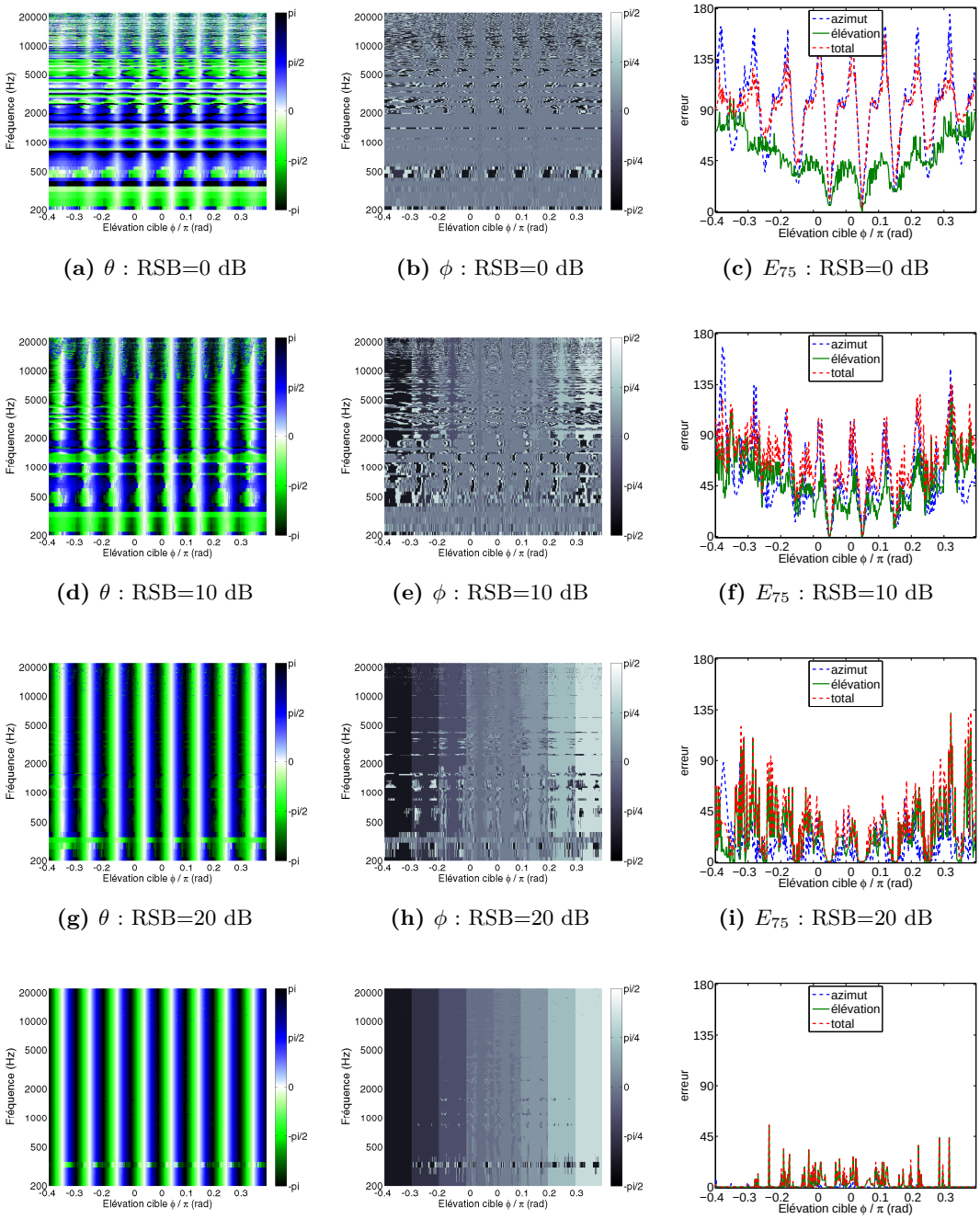
Figure V.6 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_Q, d'une source large bande suivant la trajectoire illustrée en figure IV.14 en présence d'une source de bruit avec un RSB=20 dB à des positions différentes (ligne 1 (0,0), ligne 2 $(0, \pi/2)$, ligne 3 $(0, -\pi/2)$, ligne 4 $(\pi/2, 0)$). Utilisation d'une antenne microphonique cardioïde composée de 3 capteurs pointant vers $\vec{x}$, $-\vec{x}$ et $\vec{z}$. Les capsules sur le plan horizontal sont écartées de 2 cm selon l'axe $\vec{y}$. Les estimations de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichées dans une représentation conjointe position-fréquence. Les erreurs E_{75} associées sont représentées dans la troisième colonne. Chaque ligne correspond à une direction différente de la source perturbatrice.



(a) θ **(b)** E_{75}

(c) E_θ **(d)** E_t

Figure V.7 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_Q, d'une source large bande suivant la trajectoire illustrée en figure IV.14. Utilisation d'une antenne microphonique cardioïde composée de 2 capteurs pointant vers $\vec{x}$ et $-\vec{x}$. Les capsules sont écartées de 2 cm selon l'axe $\vec{y}$. (a) Estimation de l'azimut en fonction de la fréquence et la position cible de la source. (b) Erreur d'estimation E_{75} en fonction de la position cible de la source. Erreurs d'estimation E_θ (c) et E_t (d) associées affichées dans une représentation conjointe position-fréquence.



(a) θ : RSB=0 dB **(b)** ϕ : RSB=0 dB **(c)** E_{75} : RSB=0 dB

(d) θ : RSB=10 dB **(e)** ϕ : RSB=10 dB **(f)** E_{75} : RSB=10 dB

(g) θ : RSB=20 dB **(h)** ϕ : RSB=20 dB **(i)** E_{75} : RSB=20 dB

(j) θ : RSB=30 dB **(k)** ϕ : RSB=30 dB **(l)** E_{75} : RSB=30 dB

Figure V.8 : Localisation avec la méthode des Ob-RTF, utilisant le critère de distance D_Q, d'une source large bande suivant la trajectoire illustrée en figure IV.14 en présence d'une source de bruit à (0,0) et pour différents RSB (ligne 1 0 dB, ligne 2 10 dB, ligne 3 20 dB, ligne 4 30 dB). Utilisation d'une antenne microphonique cardioïde composée de 2 capteurs pointant vers $\vec{x}$ et $-\vec{x}$. Les capsules sont écartées de 2 cm selon l'axe $\vec{y}$. Les estimations de l'azimut en colonne 1 et de l'élévation en colonne 2 sont affichés dans une représentation conjointe position-fréquence. l'erreur E_{75} associé est représentée dans la troisième colonne. Chaque ligne correspond à un RSB différent.

V.6 Conclusion

Afin de proposer une nouvelle méthode permettant la captation du son spatialisé, nous avons utilisé les caractéristiques intrinsèques des capteurs intégrés dans leur support. Ces caractéristiques étant très proches de celles des HRTF, nous avons pu utiliser les techniques de localisation existantes permettant de les exploiter. Une nouvelle méthode, appelée méthode des Ob-RTF, a été proposée. Celle-ci est basée sur les méthodes proposées par Keyruz et Mcdonalds, permettant d'une part de s'affranchir des limitations imposées par l'utilisation de la directivité des capteurs pour la localisation des sources, et, d'autre part, d'éviter une prise sonore binaurale, et par conséquent, de s'affranchir de l'utilisation d'une tête acoustique.

Différentes méthodes d'estimation de la position des sources utilisant les Ob-RTF ont été proposées. Plusieurs critères permettant d'estimer la similarité entre deux Ob-RTF (ie. distance vectorielle, Tanimoto et critère du quotient) ont été évalués. Parmi les trois critères retenus, deux permettent d'obtenir les meilleurs résultats en termes de localisation, ceci quel que soit le rapport signal à bruit (ie. Tanimoto et Quotient). Ces deux critères permettent d'atteindre des localisations des sources avec des erreurs moyennes inférieures à 16° pour un Rapport Signal à Bruit (RSB) de 30 dB, avec l'utilisation de 2 capteurs, et de 3° lorsque 3 capteurs sont utilisés. Enfin, la méthode du quotient est à privilégier, car elle permet une réduction considérable du temps de calcul par rapport à la méthode de Tanimoto.

Des études complémentaires sont à effectuer afin d'évaluer les performances de cette méthode avec des dispositifs réels, ce qui implique la mesure des Ob-RTF du dispositif microphonique.

La méthode présentée ici permet d'envisager l'utilisation d'un dispositif microphonique constitué de deux ou trois capsules microphoniques pour la localisation d'une source sonore dans l'ensemble de l'espace 3D. Elle permet également une certaine "flexibilité" du dispositif de captation, imposant au contraire une connaissance *a priori* de ce dernier et des *scénarii* d'utilisation éventuels.

VI Vers une chaîne audio 3D complète pour terminal mobile

Tout au long des cette thèse, nous avons proposé des méthodes permettant la captation et la représentation de la scène sonore 3D. Ici, nous proposons de compléter la chaîne audio jusqu'à la restitution sonore pour amener les signaux sonores 3D jusqu'aux oreilles de l'utilisateur des terminaux mobiles. Nous abordons ici les étapes manquantes, en présentant les problèmes et en proposant des pistes pour les résoudre.

VI.1 Concept général

Dans les chapitres précédents, nous avons étudié deux méthodes permettant la localisation des sources sur l'ensemble de l'espace 3D. Cette analyse directionnelle permet d'obtenir des informations nécessaires et suffisantes pour la restitution d'une scène sonore spatialisée sur un dispositif mobile, en respectant les contraintes suivantes :

- compacité du support (un signal audio et des métadonnées directionnelles),
- versatilité du choix du dispositif de restitution.

En effet, ces solutions, analogues dans leur principe au concept de SAC, permettent, par un décodage adapté, de restituer les signaux sur l'ensemble des systèmes de restitution disponibles aujourd'hui, sans perte de qualité et sans avoir recours à des ré-encodages intermédiaires.

Les deux méthodes d'analyse directionnelle reposent sur l'hypothèse qu'il n'existe qu'une seule source par bande fréquentielle à chaque trame temporelle. La longueur des trames d'analyse dépend de la résolution spectrale recherchée et de la distance maximale entre les capteurs.

Dans la phase d'analyse, les signaux sont échantillonnés à une fréquence Fe et sont analysés par trames temporelles composées de N échantillons temporels. A l'issue du procédé d'encodage directionnel, la scène sonore est composée (à chaque trame temporelle) :

- d'un signal monophonique de N échantillons,
- de N vecteurs de localisation contenant les coordonnées de directions des sources (θ_f et ϕ_f) pour chacune des N fréquences.

Compte tenu de la symétrie hermitienne de la transformée de Fourier, l'information de localisation est redondante pour les fréquences supérieures à la fréquence de Shannon. La représentation fréquentielle repose alors sur $[(N/2) - 1]$ échantillons, réduisant d'autant la taille du vecteur directionnel et allégeant par conséquence la quantité de données à stocker et à transmettre.

VI.2 Format de représentation

Le résultat du post-traitement de localisation peut être stocké de plusieurs façons. Afin de retrouver les différents éléments définissant la scène sonore, nous proposons ici une méthode de stockage des données définissant un format d'encodage.

A chaque trame, le résultat de l'encodage spatial de la scène sonore est stocké dans deux vecteurs selon une approche sans compression illustré par le tableau VI.1.

V_1 :	$s_0[1], \quad \cdots \quad s_0[N/2], \quad s_0[(N/2)+1], \quad \cdots \quad s_0[N]$	
V_2 :	$\theta[f_0], \quad \cdots \quad \theta[Fs/2]$	$\phi[f_0], \quad \cdots \quad \phi[Fs/2]$

Tableau VI.1 : Représentation de la scène sonore encodée.

Le premier vecteur (V_1) contient les N échantillons du signal temporel $s_0[n]$ représentatif de la scène sonore. Le second vecteur (V_2) comprend les directions θ et ϕ pour chacune des fréquences f. Le vecteur $V2$ est composé de deux parties : la première contient les valeurs de θ et la seconde les valeurs de ϕ pour des fréquences allant de 0 Hz à la fréquence de Shannon.

Cette approche permet de garantir une synchronisation entre le signal sonore $s_0[n]$ et la direction précise des sources. L'encodage des valeurs de direction sur uniquement

6 bits permet une résolution de 5,6 ° en azimut et de 2,8 ° en élévation, en considérant le repère utilisé dans ce document (figure IV.1). Cet encodage réduit considérablement la quantité de l'information (3 :1), tout en restant compatible avec le flou de localisation de ± 3°[Blauert, 1983] (I.2.4) et conforme aux travaux de Daniel [Daniel, 2011].

VI.3 Restitution

Dans la perspective d'une utilisation en mobilité, le binaural a les atouts nécessaires en termes de compacité, légèreté et préservation de l'intimité (II.1.1.b).

La restitution binaurale est effectuée par un filtrage de la source sonore avec la HRIR ou la *Binaural Room Impulse Response* (BRIR) correspondant à la direction de la source (cf. section I.4.2).

VI.3.1 Décodage binaural

La synthèse binaurale est basée sur le produit de convolution entre un signal sonore, généralement enregistré dans un milieu anechoïque, et la HRIR des deux oreilles correspondant à la direction où l'on souhaite placer virtuellement la source.

Dans le domaine fréquentiel, la synthèse binaurale s'écrit sous la forme

$$\begin{bmatrix} S_L(f) \\ S_R(f) \end{bmatrix} = \begin{bmatrix} HRTF_L(\theta,\phi)(f) \\ HRTF_R(\theta,\phi)(f) \end{bmatrix} S(f), \tag{VI.1}$$

où $S(f)$ représente le signal à placer dans l'espace, et $S_L(f)$ et $S_R(f)$ les signaux binauralisés dans la direction (θ,ϕ) respectivement pour les oreilles gauche et droite.

Cette représentation est compatible avec l'encodage présenté aux chapitres précédents, car on dispose des directions pour chaque fréquence d'un signal $S_o(f)$ représentatif de la scène sonore (tableau VI.1). Il suffit donc d'appliquer un gain fréquentiel, à chaque source identifiée, correspondant aux HRTF associées à sa direction.

A chaque trame de N échantillons, le signal $s_0[n]$ contenu dans le vecteur V_1 est passé dans le domaine fréquentiel par *Fast Fourier Transform Function* (FFT) et devient $S_0(f_k)$ (avec $f_k = kF_e/N$, N étant le nombre de points de la FFT, et $k \in [0, N/2-1]$). Les directions des sources associées à chaque fréquence permettent d'identifier les HRTF correspondant aux directions des sources souhaitées $(\hat{\theta},\hat{\phi})$ parmi les directions des HRTF disponibles.

A cet instant, deux cas sont possibles : soit la direction $(\hat{\theta}, \hat{\phi})$ fait partie ou ne fait pas partie du jeu d'HRTF utilisé. Dans le cas où elle n'en fait pas partie, plusieurs solutions sont possibles :

- soit l'HRTF la plus proche de la direction souhaitée est utilisée,
- soit le groupe des HRTF les plus proches est utilisé en leur appliquant une pondération prenant en compte leur distance à la direction souhaitée [Pulkki, 2002],
- soit le jeu d'HRTF est interpolé dans la direction souhaitée à partir des HRTF disponibles (I.4.2.e).

A chaque fréquence f_k, le signal $S_o(f_k)$ est ensuite multiplié par les valeurs de $HRTF_{l,r}(\hat{\theta}, \hat{\phi})(f_k)$ relatives à la direction $(\hat{\theta}, \hat{\phi})$ selon l'équation (VI.1).

Les signaux $S_L(f_k)$ et $S_R(f_k)$ sont dupliqués afin de compléter les fréquences supérieures à la fréquence de Shannon par symétrie hermitienne. Finalement, une FFT inverse est appliquée aux signaux afin de retourner dans le domaine temporel.

L'utilisation d'une fenêtre de pondération temporelle à transition douce permet une meilleure décomposition fréquentielle du signal temporel $s_0[n]$. Afin de compenser la perte d'énergie engendrée par la fenêtre, un recouvrement temporel des trames est nécessaire. Cette opération effectue un lissage temporel de la localisation des sources évitant ainsi les sauts de position.

VI.3.2 Suivi des mouvements de tête

Si aucune correction n'est apportée, la scène auditive restituée au casque suit les déplacements de la tête de l'auditeur, ce qui dégrade l'immersion sonore. Il est donc nécessaire de mesurer ces mouvements (*head tracking*) afin de les compenser. Cette correction permet, d'une part, d'offrir une restitution sonore conforme à une situation réaliste, et d'autre part, de donner au système auditif des indices complémentaires (indices dynamiques de localisation), permettant de stabiliser la position de la source sonore (I.2.3, I.4.2.d). Aujourd'hui, la plupart des terminaux mobiles sont équipés de dispositifs permettant la détection des mouvements et d'orientation, tels que des gyroscopes, accéléromètres, boussoles numériques et GPS [Lane et al., 2010]. L'utilisation de ces capteurs permettrait alors d'estimer la position de l'auditeur afin de mettre en œuvre le *head-tracking*.

D'une part, cette information permet d'effectuer une correction d'une scène sonore enregistrée préalablement ou en transmission depuis un autre dispositif afin de corriger les mouvements de l'auditeur. D'autre part, les informations de géolocalisation permettent aussi contrôler le déclenchement de contenus produisant ainsi une réalité augmentée adaptée au lieu d'écoute.

Nous décrivons ici une méthode de mise en œuvre du *head-tracking* compatible avec les méthodologies décrites dans les chapitres IV et V et avec le décodage binaural décrit précédemment.

Les données de rotation issues des capteurs de position peuvent être exprimées comme des rotations autour des trois axes en coordonnées cartésiennes $(\vec{x}, \vec{y}, \vec{z})$ et reçoivent respectivement les noms de (*roll*, *pitch*, *yaw*) ou (*tilt*, *tumble*, *rotate*). Ces rotations sont exprimées respectivement comme des valeurs d'angles (α_h, β_h, γ_h), où l'indice h indique qu'il s'agit de la rotation de la tête.

Les matrices de rotation autour de chaque axe sont

$$R_x(\alpha_h) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_h & -\sin\alpha_h \\ 0 & \sin\alpha_h & \cos\alpha_h \end{pmatrix}_{\mathscr{B}_c}, \tag{VI.2a}$$

$$R_y(\beta_h) = \begin{pmatrix} \cos\beta_h & 0 & \sin\beta_h \\ 0 & 1 & 0 \\ -\sin\beta_h & 0 & \cos\beta_h \end{pmatrix}_{\mathscr{B}_c}, \tag{VI.2b}$$

$$R_z(\gamma_h) = \begin{pmatrix} \cos\gamma_h & -\sin\gamma_h & 0 \\ \sin\gamma_h & \cos\gamma_h & 0 \\ 0 & 0 & 1 \end{pmatrix}_{\mathscr{B}_c}, \tag{VI.2c}$$

et la matrice de rotation totale R est définie par le produit matriciel

$$R = R_x(\alpha_h) R_y(\beta_h) R_z(\gamma_h). \tag{VI.3}$$

La correction du signal sonore se fait en appliquant les rotations inverses $-\alpha_h$, $-\beta_h$ et $-\gamma_h$ aux vecteurs de localisation des sources issues du procédé d'encodage. Le produit matriciel n'étant pas commutatif, l'ordre du produit doit être également inversé pour obtenir la matrice de correction $\hat{R}$ suivant

$$\hat{R}(\alpha_h, \beta_h, \gamma_h) = R_z(-\gamma_h) R_y(-\beta_h) R_x(-\alpha_h). \tag{VI.4}$$

De cette façon, on détermine les directions des HRTF à prendre en compte pour la restitution ou la synthèse des signaux. Les signaux sont ainsi restitués dans un repère global qui ne dépend plus de la position de la tête de l'auditeur.

VI.3.3 Autres modes de restitution

Comme nous l'avons évoqué en I.5.1, les méthodes d'encodage de type "format objet" permettent la restitution d'une scène sonore ainsi encodée sur tout type de dispositif de restitution. La méthode présentée ici est analogue à ce type de format a condition d'effectuer un suivi des sources composant la scène sonore et peut, au même titre, être implémentée sur tout dispositif multi-haut-parleur.

Dans un ordre croissant de haut-parleurs disponibles, nous pouvons citer les techniques suivantes.

1. **Binaural :** le rendu binaural sur haut-parleur est effectué grâce aux techniques comme le *transaural* et le stéréo dipôle (I.4.2.d). Ces méthodes permettent l'adaptation de signaux binauraux obtenus après une première phase de décodage comme celle présentée en VI.3.
2. **Stéréo multicanal :** le rendu stéréophonique multicanal est possible en utilisant des techniques comme le *Vector Based Amplitude Panning* (VBAP) [Pulkki, 2002]. Elle permet d'effectuer un panoramique d'intensité entre les canaux disponibles. L'amplitude des signaux délivrés par les différents haut-parleurs est pilotée par les vecteurs déterminant la position des sources, dans le but de créer des sources fantômes entre les haut-parleurs disponibles (I.3.1.b).
3. **Ambisonique :** la restitution sur un dispositif ambisonique est également possible. La position des sources obtenue dans la phase de localisation détermine alors le ré-encodage de la scène sonore dans une base d'harmoniques sphériques, où l'ordre de l'ambisonique est directement lié au nombre de haut-parleurs disponibles pour la restitution (I.4.3.b).
4. **WFS :** le décodage peut être effectué sur un dispositif WFS où les sources sont disposées par un contrôle des différences d'amplitude et de temps entre les haut-parleurs disponibles pour synthétiser un front d'onde dans la direction désirée.

VI.4 Plus loin sur la prise de son en mobilité : mise en œuvre du hand-tracking

Par analogie au *head-tracking*, nous proposons le *hand-tracking* correspondant au "suivi de la main", ou plus exactement à l'ensemble de procédés permettant de connaître la position des microphones lors de la captation. Les données ainsi stockées sont utilisées dans une phase de post-traitement de la scène sonore.

Un système de prise de son intégré à un dispositif mobile est généralement posé sur une surface telle qu'une table, ou tenu à la main. Dans les deux cas, l'utilisation des capteurs de position permet de déterminer la position du dispositif par rapport à un repère global. Il est ainsi possible de présenter, au moment de la restitution, les sources dans une position analogue à celle où elles se trouvaient lors de la prise sonore, en s'affranchissant des éventuels déplacements de l'enregistreur. Dans le premier cas, la correction permet une représentation correcte de la scène sonore et dans le second, des rotations éventuellement gênantes engendrées par les déplacements de l'utilisateur au cours de l'enregistrement sont également corrigées.

Si les paramètres de position du dispositif sont stockés tout au long de l'enregistrement, il est ensuite possible d'effectuer une correction des mouvements de ce dernier de la même manière qu'il est possible de corriger les mouvements de la tête. Cette correction peut être effectuée lors de l'encodage directionnel de la scène sonore ou au moment de la restitution.

Dans le premier cas, le stockage des paramètres de position du capteur est inutile, à moins que l'opération inverse ne soit envisagée. Si la correction est effectuée au moment de la restitution, se pose alors le problème du stockage de ces paramètres, car ils interviennent directement dans le volume total de données à stocker et/ou à transmettre. Dans un premier temps, nous recommandons de stocker les informations de position sous forme scalable[1] : d'une part, la position globale du capteur à une fréquence d'échantillonnage inférieure à 1 Hz, et d'autre part une information complémentaire par le stockage des déplacements relatifs avec une fréquence d'échantillonnage supérieure.

[1] La fréquence d'échantillonage est adaptée à la vitesse des variations

VI.5 Conclusion

Les méthodes présentées ici permettent d'effectuer la transmission et la restitution de la scène sonore, complétant ainsi la chaîne électroacoustique 3D dédiée aux terminaux mobiles.

Les signaux résultant de l'analyse de la scène sonore par localisation des sources permettent une compression considérable des données à stocker et à transmettre. La méthode proposée ici permet en effet une compression de 3 :1 dans une approche simplifiée. L'utilisation de l'identification et de la séparation des sources, ainsi que l'introduction dans l'analyse de phénomènes perceptifs, tels que le masquage temporel, énergétique et directionnel, permettraient de réduire d'avantage les paramètres de représentation de la scène sonore, pour ne conserver que le minimum d'informations nécessaires pour sa perception lors de la restitution.

A la fin de la chaîne se trouve l'auditeur. Il sera donc nécessaire d'évaluer l'impact de cette représentation de la scène sonore, ainsi que l'impact de la compression apportée sur la qualité de l'expérience perçue.

Nous avons présenté comment les différents capteurs de mouvement intégrés dans les terminaux mobiles permettent d'effectuer un suivi de la tête afin d'améliorer le rendu binaural. Les terminaux mobiles sont utilisés le plus fréquemment en mobilité et dans des ambiances acoustiques souvent bruyantes. Reste à envisager la possibilité d'utiliser les microphones intégrés dans les terminaux mobiles au moment de la restitution sonore, dans deux applications. La première consiste à fournir un contrôle actif du bruit perçu par l'auditeur améliorant ainsi le confort d'écoute. La deuxième permet de profiter du masquage du bruit ambiant pour simplifier le rendu sonore, car, en présence des sources multiples, le flou de localisation s'élargit, permettant un rendu moins précis de la position des sources [Daniel, 2011].

Conclusions et perspectives

Conclusions

Ce document a présenté l'ensemble des travaux que j'ai effectués pendant mon doctorat au sein de l'équipe *son 3D* à *Orange Labs* sous la co-direction de Rozenn Nicol ingénieure de recherche et chef de projet, et de Laurent Simon, enseignant chercheur, dans le cadre d'un partenariat avec le Laboratoire d'Acoustique de l'université du Maine (LAUM). Il s'agit de travaux de recherche visant à intégrer des outils de son spatialisé dans les terminaux mobiles en prenant en compte les contraintes imposées d'une part, par les techniques liées à la spatialisation sonore et d'autre part, par les dispositifs mobiles.

Nous avons tout d'abord effectué une description des différentes techniques de son 3D et présenté les spécificités des dispositifs mobiles, du point de vue technologique et de leur usage. A l'issue de cette première étape, nous avons pu effectuer une mise en parallèle de ces deux technologies afin de définir leur terrain commun, et nous avons détaillé les technologies de spatialisation sonore pouvant être adaptées à une utilisation en mobilité.

La restitution binaurale, de par sa compacité et sa légèreté (reproduction sur casque stéréo), s'est avérée la méthode de restitution privilégiée pour une utilisation en mobilité. La prise de son binaurale comportant un grand nombre d'inconvénients, nous avons cherché des solutions alternatives pour l'étape de captation.

La mise en place d'une première maquette, permettant de simuler une chaîne audio 3D complète pour diffusion au casque, a démontré que la qualité perçue de la captation d'une scène sonore pour une restitution binaurale dépend tout autant du post-traitement utilisé que du dispositif de captation.

Suite à ce constat, nous avons orienté les recherches vers la conception d'un dispositif microphonique permettant d'effectuer une *Auditory Scene Analysis* (ASA) de forme simple, afin d'atteindre une représentation de la scène sonore analogue au "format objet".

Nous avons alors défini et évalué une configuration microphonique utilisant uniquement trois capteurs pour l'estimation de direction dans tout l'espace 3D. Ce dispositif est couplé à un post-traitement spatial pour effectuer l'*Auditory Scene Analysis* (ASA). Deux méthodes d'ASA ont été mises au point. La première utilise les directivités des capteurs et la deuxième, appelée Ob-RTF par analogie aux HRTF, profite des modifications directionnelles induites par le positionnement des microphones près d'un corps diffractant. Ces techniques ont montré leur efficacité dans des configuration idéales et en présence de sources perturbatrices.

La chaîne sonore a été complétée en définissant un format de stockage et/ou de transport permettant le décodage sur tout type de système de restitution, notamment pour une écoute binaurale. Il a été également montré comment les dispositifs de localisation intégrés dans les terminaux mobiles peuvent être utilisés pour l'amélioration de la captation et du rendu sonore.

Perspectives

Les évaluations des dispositifs microphoniques présentées ici ont été effectuées principalement sur une base de simulations numériques. La prochaine étape est leur évaluation à partir de dispositifs réels, tout d'abord dans des environnements anéchoïques, avant de poursuivre leur analyse dans des configurations acoustiques plus réalistes.

Les algorithmes de localisation développés ici, ayant pour objectif de démontrer la faisabilité et d'analyser les performances des techniques, ont besoin d'être optimisés pour réduire leur complexité et les rendre compatibles avec les contraintes de calcul des terminaux mobiles. A ce titre, un prototype intégrant les technologies que nous avons développées pendant la présente thèse est en cours de développement à Orange Labs sous un environnement iOS [Magnoux and Lefort, 2014, Nujsop, 2013].

Dans cette étude, nous avons négligé plusieurs aspects perceptifs. En intégrant des méthodologies de séparation et d'identification des sources, ainsi que des phénomènes de masquage (énergétique, spectral, temporel et spatial), les performances des techniques développées pourraient être améliorées en termes d'efficacité sans dégrader la qualité du rendu sonore. La mise en place de ce type de méthodes doit être effectuée avec soin, afin de ne pas augmenter la complexité algorithmique aux dépens du rendu sonore. De même, les méthodes d'évaluation des performances proposées doivent être revisitées, car elles ont pour objectif de comparer des positions cibles des sources avec les positions obtenues après traitement. La prise en compte de la perception, n'obligeant plus cette correspondance, impose la modification des critères d'évaluation qui doivent s'appuyer

sur des tests d'écoute.

La connectivité des terminaux mobiles a aussi également été négligée. Cette caractéristique pourrait permettre de s'affranchir des limitations de capacité de calcul des terminaux mobiles, en déportant les traitements et le stockage des données audio sur des serveurs dédiés, dans une approche de type *cloud computing*. Dans ce cas, des nouvelles stratégies de compression et de transmission des données doivent être développées afin de ne pas remplacer le problème de complexité de calcul par une saturation des dispositifs de transmission, également très gourmands en termes de consommation énergétique.

L'intégration du son spatialisé sur des terminaux mobiles est proche d'être une réalité mais il reste toujours aujourd'hui des difficultés à surmonter, pour lesquelles nous espérons avoir contribué en posant ici les premières pierres de cette intégration.

A Localisation des sources

Les méthodes d'encodage directionnel présentées en I.5.2.b, reposent sur une première phase d'analyse de la scène sonore, permettant la localisation des sources à partir des signaux ambisoniques à l'ordre 1. Ces méthodes peuvent être considérées comme un encodage spatial de la scène sonore pouvant aboutir dans un format objet.

Cette phase de localisation étant un pré-requis pour l'encodage d'une scène sonore spatiale réelle, il est intéressant de présenter les principales méthodes de localisation de sources.

A.0.1 Analyse de la scène sonore auditive ASA

La localisation des sources sonores, effectuée dans le cadre de l'analyse de scènes sonores *Auditory Scene Analysis* (ASA), prend en compte les mécanismes de localisation sonore mis en jeu par le système auditif. L'*Interaural Time Difference* (ITD) et la différence interaurale d'intensité *Interaural Intensity Difference* (IID) (correspondant à l'*Interaural Level Difference* (ILD)) sont utilisés, et des approches extraites de la psychologie de la forme, ou "Gestaltisme", sont généralement prises en compte [Bregman, 1994]. Ainsi, Bregman montre que :

- deux événements acoustiques qui commencent et se terminent au même temps sont perçus comme un événement unique,
- la progression de la transformation au cours d'un événement sonore est lente s'il s'agit d'un événement isolé et que la transformation d'une série événements isolés est lente s'ils sont générés par la même source,
- les oscillations ayant une fréquence multiple d'une composante plus grave appartiennent souvent à la même source (cette condition est maintenue même s'ils surviennent au cours de l'événement sonore, à condition que la fréquence fondamentale soit toujours présente),

- les modifications d'un signal sonore affectent toutes les composantes du son résultant de façon identique et simultanée.

Potentiellement, ces paramètres peuvent être utilisés pour regrouper les sources suite à une analyse spectrale comportant des modèles auditifs, car une scène sonore est rarement composée d'une seule source statique et continue.

A.0.2 Méthodes de localisation des sources

Une prise sonore permet de connaître l'information rélative à la pression acoustique à un instant donné à l'emplacement du capteur. Cette information étant dépendante de la source sonore et de sa position dans l'espace par rapport au capteur, il est possible de décrire la source et sa position à partir des signaux sonores captés.

Pour obtenir la position d'une source, il est nécessaire de traiter et comparer les signaux sonores entre eux, afin d'en extraire les informations suffisantes pour remonter à une coordonnée physique relative à la position de la source [Burdic, 1991].

Lorsqu'un microphone unique est utilisé, il n'est pas possible d'obtenir des informations permettant d'atteindre une coordonnée spatiale. Lorsqu'au moins deux capteurs sont utilisés, deux informations peuvent en être extraites. La première porte sur la différence de temps entre les deux signaux et la seconde sur leur différence d'intensité. Spiesberger [Spiesberger, 2001] montre que la localisation d'une source est possible grâce à l'utilisation du retard du signal entre les microphones et que le nombre de microphones doit être déterminé en fonction du nombre de dimensions à estimer (n+1 microphones pour une localisation sur n dimensions).

Il est possible de représenter la scène sonore composée de N sources et captée par une antenne microphonique de L capteurs par l'expression

$$\mathbf{x}(t,\omega) = \sum_{n=1}^{N} \mathbf{d}(t,\tau_n) S_n(t,\omega) + \mathbf{b}(t,\omega)\ , \tag{A.1}$$

où $\mathbf{x}(t,\omega)$ est la matrice des signaux microphoniques captés à l'instant t et à la pulsation ω, par une antenne microphonique de vecteur directionnel $\mathbf{d}(t,\tau_n)$, et générée par l'ensemble des sources représentées par $S_n(t,\omega)$ et soumise à un bruit $\mathbf{b}(t,\omega)$ représentant les sources parasites, les réverbérations et les échos.

Le vecteur $\mathbf{d}(t, \tau_n)$ est défini par

$$\mathbf{d}(t, \tau_n) S_n(t, \omega) = \begin{pmatrix} \mathcal{G}_{1n} e^{-i\omega\tau_{1n}} \\ \vdots \\ \mathcal{G}_{ln} e^{-i\omega\tau_{ln}} \\ \vdots \\ \mathcal{G}_{Ln} e^{-i\omega\tau_{Ln}} \end{pmatrix} , \tag{A.2}$$

où $\mathcal{G}_{ln}$ et τ_{ln} correspondent respectivement au gain et au retard induits à la source n par le microphone l, par rapport à un gain et un retard de référence choisis arbitrairement. Par exemple, si le microphone de référence est le microphone 1, l'équation A.2 devient

$$\mathbf{d}(t, \tau_n) S_n(t, \omega) = \begin{pmatrix} 1 \\ \vdots \\ \mathcal{G}_{ln} e^{-i\omega\tau_{ln}} \\ \vdots \\ \mathcal{G}_{Ln} e^{-i\omega\tau_{Ln}} \end{pmatrix} , \tag{A.3}$$

les deux facteurs $\mathcal{G}_{ln}$ et τ_{ln} dépendant directement de la position de la source, de l'emplacement des capteurs et de leur directivité.

Généralement, la source est supposée suffisamment éloignée des capteurs pour approximer τ suivant,

$$\tau \approx \frac{d}{c} \cos\theta \, , \tag{A.4}$$

où d est la distance entre les microphones et θ l'angle de la source par rapport à la droite colinéaire à la distribution des microphones.

Pour la plupart des applications, les capteurs sont considérés comme étant omnidirectionnels, ce qui permet d'approcher les termes $\mathcal{G}_{ln}, \forall n$, par

$$\mathcal{G}_{ln} \approx 1 \, . \tag{A.5}$$

Considérant ces hypothèses, la localisation des sources revient à effectuer une analyse de la différence de temps d'arrivée des signaux entre les différents capteurs *Time Difference Of Arrival* (TDOA).

Aujourd'hui, la plupart des méthodes agissent dans le domaine fréquentiel [Blandin et al., 2011a]. Elles partent de l'hypothèse qu'il n'existe qu'une seule source à chaque fréquence par trame temporelle, ou, de manière équivalente, elles déterminent une source unique sur l'ensemble du spectre à chaque trame temporelle.

Il est possible de classifier les algorithmes de localisation en trois grandes familles [Blandin et al., 2011a]. La première effectue une recherche de pics sur l'histogramme des temps d'arrivée dans chaque échantillon temps-fréquence [Faller and Merimaa, 2004, Viste and Evangelista, 2003]. La deuxième utilise la phase du spectre [Knapp and Carter, 1976] et la troisième se sert des principes de regroupement ou *clustering* en anglais [Sawada et al., 2007].

La première famille est limitée à des dispositifs comportant un espacement faible pour éviter le repliement spatial [Sawada et al., 2007]. La deuxième, originaire de la communauté de l'antennerie microphonique, est basée principalement sur l'analyse du TDOA pour l'estimation des directions des sources, se basant sur des pics fréquentiels à partir d'un certain seuil [Schmidt, 1986]. La troisième peut être utilisée pour toute configuration microphonique, et est très sensible aux paramètres d'initialisation des algorithmes définissant l'écartement minimum entre chaque *cluster* [Sawada et al., 2007].

La première famille, étant basée sur l'ITD décrit en I.2.1, elle ne présente pas de difficulté particulière. Nous allons approfondir la description des deux autres familles.

A.0.3 Méthodes basées sur le spectre de phase

Ces méthodes reposent sur l'estimation d'un spectre de phase désigné par une fonction $\Phi(\tau)$ dépendant du TDOA τ [Blandin et al., 2011b].

L'utilisation de ces méthodes pour la localisation d'une source implique un repliement spatial. Lorsque la distance entre les capteurs est supérieure à la longueur d'onde, il n'est plus possible de différencier la phase obtenue entre les signaux captés par une paire microphonique. Pour résoudre ce problème, l'intégration temporelle et/ou fréquentielle est souvent effectuée. Cette intégration est souvent réalisée par sommation du spectre ou par la recherche du maximum de celui-ci.

Les méthodes employées pour la localisation des sources par différence de temps entre les capteurs utilisent l'inter-corrélation des signaux microphoniques ou sa généralisation dans le domaine fréquentiel, la méthode ***Generalized Cross-Correlation*** **(GCC)**.

Cette méthode introduite par Knapp [Knapp and Carter, 1976] utilise une fonction

sinusoïdale pour le calcul du spectre de phase local dont l'amplitude dépend des signaux mis en jeu et est communément connue sous le nom GCC-PHAT.

Le spectre de phase obtenu avec cette méthode est noté $\Phi^{GCC-PHAT}$. Il est défini pour deux signaux $x_1(t)$ et $x_2(t)$ et l'analyse est effectuée sur des trames temporelles en normalisant le spectre [Knapp and Carter, 1976] grâce à la relation

$$\Phi^{GCC-PHAT}(t,\omega,\theta,\phi) = \Re\left(\frac{X_1(t,\omega)X_2^*(t,\omega)}{|X_1(t,\omega)X_2^*(t,\omega)|}e^{-i\omega\tau(\theta,\phi)}\right) , \tag{A.6}$$

où $X_1(t,\omega)$ et $X_2(t,\omega)$ sont les spectres de $x_1(t)$ et $x_2(t)$ à l'instant t et la pulsation ω, et où $*$ représente ici la matrice transconjuguée.

La méthode ***MUltiple Signal Classification*** **(MUSIC)** [Schmidt, 1986] est une variante de la méthode GCC-PHAT. Elle utilise la matrice de covariance du signal calculé au voisinage fréquentiel de la fenêtre W, défini comme

$$\hat{\mathbf{R}}_{XX}(t,\omega) = \frac{\sum_{t',\omega'} \mathrm{W}(t'-t,\omega'-\omega)X(t',\omega')X(t',\omega')^*}{\sum_{t',\omega'} \mathrm{W}(t'-t,\omega'-\omega)} . \tag{A.7}$$

Elle cherche à faire correspondre le vecteur directionnel $\mathrm{d}(\omega,\theta,\phi)$ défini dans (A.3) avec le premier vecteur principal $\mathrm{v}(t,\omega)$ de la matrice de covariance $\hat{\mathbf{R}}_{XX}(t,\omega)$, suivant

$$\Phi^{MUSIC}(t,\omega,\theta,\phi) = \left(1 - \frac{1}{2}|\mathrm{d}(\omega,\theta,\phi)^*\mathrm{v}(t,\omega)|^2\right)^{-1} . \tag{A.8}$$

Nesta [Nesta et al., 2009] propose, dans une autre approche, d'identifier par Analyse par Composantes Principales (ACP) deux sources par trame tempo-fréquentielle. Le spectre de phase contenant ces deux sources est exprimé en fonction des TDOA correspondants (τ_1 et τ_2) et des coefficients de pondération $r_1(t,\omega)$ $r_2(t,\omega)$. Cette méthode appelée "transformée de l'état de cohérence cumulative" ***cumulative State Coherence Transform*** **(cSCT)**, définit le spectre de phase par

$$\Phi^{cSCT}(t,\omega,\theta,\phi) = \sum_{n=1}^{2} \rho\left[\frac{1}{2}\left|e^{-i\omega\tau_n} - r_n(t,\omega)\right|\right] , \tag{A.9}$$

avec $\rho(u) = 1 - \tanh(\alpha\sqrt{u})$, où α est une constante à déterminer.

Ces méthodes donnant la même importance à toutes les composantes fréquentielles, Blandin [Blandin et al., 2011b] a proposé d'attribuer un poids à chaque composante fréquentielle en fonction du Rapport Signal-à-Bruit (RSB). La pondération spectrale peut être extraite de la fonction de cohérence $\gamma^{coh}(t,\omega)$

$$\gamma^{coh}(t,\omega) = \frac{RSB}{1+RSB} = \frac{|R_{X_1X_2}(t,\omega)|}{\sqrt{R_{X_1X_1}(t,\omega)R_{X_2X_2}(t,\omega)}}\,, \tag{A.10}$$

et la relation (A.6) devient

$$\Phi^{GCC-PHAT}(t,\omega,\theta,\phi) = \Re\left(\frac{X_1(t,\omega)X_2^*(t,\omega)}{|X_1(t,\omega)X_2^*(t,\omega)|}\frac{{\gamma^{coh}}^2(t,\omega)}{1-{\gamma^{coh}}^2(t,\omega)}\cdot e^{-i\omega\tau(\theta,\phi)}\right)\,. \tag{A.11}$$

Valin et al. proposent [Valin et al., 2007] une méthode de pondération différente, repérant les moments de silence avec une méthode du contrôle récursif du minimum moyen, *Minima Controlled Recursive Averaging* (MCRA), où à chaque récurrence i

$$\gamma_i^{MCRA}(t,\omega) = \frac{RSB_i}{1+RSB_i}\,, \tag{A.12}$$

et la relation (A.6) devient, pour deux récurrences consécutives,

$$\Phi^{GCC-MCRA}(t,\omega,\theta,\phi) = \Re\left(\frac{X_1(t,\omega)X_2^*(t,\omega)}{|X_1(t,\omega)X_2^*(t,\omega)|}\gamma_1^{MCRA}(t,\omega)\gamma_2^{MCRA}(t,\omega)e^{-i\omega\tau(\theta,\phi)}\right)\,. \tag{A.13}$$

A.0.4 Méthodes de regroupement ou *clustering*

Les méthodes présentées jusqu'à maintenant ne tiennent pas compte du RSB pour la recherche des TDOA. Les méthodes de regroupement se basent sur cette mesure pour une première détermination des TDOA. Cette phase d'initialisation permet ainsi de déterminer les groupes.

La méthode de Sawada [Sawada et al., 2007] propose un regroupement des sources par

rapport à une distance euclidienne pour ensuite recalculer le TDOA pour chaque groupe.

Il est également possible d'effectuer le regroupement à partir d'une approche statistique. Izumi[Izumi et al., 2007] propose par exemple d'effectuer un algorithme d'espérance-maximisation à partir de l'hypothèse que chaque trame temps-fréquence (t, ω) contient une source prédominante $n_{t\omega}$ et un bruit diffus,

$$x(t,\omega) = s_{n_{t\omega}}(t,\omega)d(\omega,\tau_{nt\omega}) + b(t,\omega)\ . \tag{A.14}$$

Les paramètres de la source $s_{n_{t\omega}}(t,\omega)$ sont considérés comme déterministes et le bruit $b(t,\omega)$ est considéré comme un vecteur aléatoire respectant une loi gaussienne. La matrice de covariance associée à ce dernier est $v^b\Psi(\omega)$ où v^b est une constante et où

$$\Psi(\omega) = \begin{pmatrix} 1 & \mathrm{sinc}(\omega\frac{d}{c}) \\ \mathrm{sinc}(\omega\frac{d}{c}) & 1 \end{pmatrix}\ , \tag{A.15}$$

où d la distance entre deux capteurs, c la célérité du son et $\mathrm{sinc}(u) = \frac{\sin(u)}{u}$. D'autres variantes ont été proposées en modifiant l'algorithme d'espérance-maximisation [Blandin et al., 2011a], permettant ainsi de détecter plusieurs sources par trame temps-fréquence.

Ces algorithmes ne prennent pas compte la directivité des capteurs. Pour pallier ce problème, des méthodes permettant de les compenser ont été proposées [Ba et al., 2007]. Nous considérons que pour des applications audio, cette caractéristique est une composante importante qui mérite d'être exploitée.

B État de l'art dans le domaine des microphones de silicium ou MEMS

B.1 Introduction

Compte tenu de l'évolution de la production de systèmes audiovisuels à taille réduite tels que les téléphones mobiles, consoles de jeu, ordinateurs portables, entre autres, les constructeurs ont un besoin croissant de miniaturiser l'ensemble des composants. Le microphone est l'un des composants qui est toujours présent dans ce type d'appareils. Dans un souci de taille et de réduction des coûts de production, les fabricants ont cherché à produire des microphones utilisant les mêmes méthodes employées dans la fabrication des composantes de silicium. Le premier prototype fabriqué utilisant cette technologie a été réalisé en 1983 [Scheeper et al., 1994] par Royer [Royer et al., 1983]. Aujourd'hui, plusieurs fabricants commercialisent déjà des microphones intégrés dans une puce électronique, ce qui leur a valu le nom de microphones de silicium.

B.2 Technologie de fabrication

Actuellement la miniaturisation des composants électroniques est basée sur des systèmes dites *Micro-electro-mechanical Systems* (MEMS). Cette technologie permet de fabriquer des composants électromécaniques de taille variant entre 1 et 100 μm pour constituer un ensemble pouvant aller de 20 μm à 1 mm. Les avantages de ce type de dispositif sont notamment, la résistance et la durabilité du silicium et la facilité de fabrication, par des procédés utilisés dans la fabrication de composants électroniques tels que la photolithographie et le grattage. Les procédés utilisés présentent également l'avantage

de pouvoir produire en masse des dispositifs présentant très peu de variations des performances. En outre, cette technologie a permis d'intégrer directement l'amplificateur dans la même plaque de silicium, réduisant ainsi les distances de câblage et les risques de mauvais contacts lors du montage tout en réduisant la sensibilité aux bruits parasites. Actuellement, certains fabricants comme *Akustika, Analog-devices, Infineon* ente autres, commercialisent des microphones intégrant directement un Convertisseur Analogique - Nunérique (CAN) assurant une immunité aux perturbations électromagnétiques. Le principal inconvénient des dispositifs de petite taille est l'impossibilité de l'évacuation de la chaleur qui se traduit par un niveau de bruit très important en sortie des capteurs.

B.3 Types de microphones

Un microphone est un dispositif qui assure la transduction de l'énergie acoustique en énergie électrique de manière directe ou indirecte. Pour la fabrication des microphones de silicium les développeurs ont voulu adopter les différents principes de transduction utilisés pour la fabrication des microphones traditionnels. Le transfert **acoustique – mécanique** est toujours assuré par la mise en vibration d'une membrane tendue. Les différents auteurs ont proposé différentes géométries de membrane, octogonales [Scheeper et al., 2003], rectangulaires [Tajima et al., 2005], carrées [Goto et al., 2007], circulaires [Hall et al., 2008a] ou de formes spécifiques [Fuldner et al., 2005].

Après les premiers essais, les microphones ont montré une faiblesse en matière de sensibilité. Dans le but d'augmenter la sensibilité, la souplesse de la membrane a été augmentée. Certains auteurs ont proposé la mise en place de ressorts et Füldner [Fuldner et al., 2005] a réussi à augmenter la sensibilité de 1,2 mV/Pa pour une membrane plate à 8,2 mV/Pa en creusant huit nervures sur son pourtour. La transduction électroacoustique détermine actuellement le type de transducteur. Actuellement, nous pouvons dénombrer quatre types de transduction :

- Microphones capacitifs
- Microphones piézoélectriques
- Microphones optiques
- Microphone FET (Transistor à effet de champ)

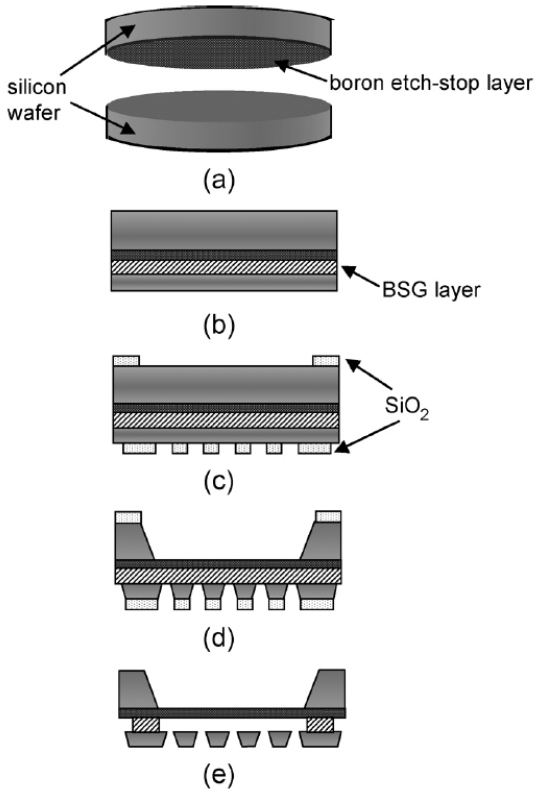


Figure B.1 : Procédé de fabrication d'un microphone en silicium à condensateur extrait de [Goto et al., 2007].

B.3.1 Microphones capacitifs

Il s'agit du type le plus répandu de microphones dû à leur facilité de mise en œuvre et leurs performances acoustiques. Leur fonctionnement est identique à celui des microphones capacitifs traditionnels. Il s'agit de deux électrodes séparées par une lame d'air constituant ainsi un condensateur plan. L'une des électrodes est fixe (contre-électrode) et l'autre est mobile et constitue la membrane du microphone.

La capacité C est déterminée par la relation,

$$C = \epsilon \frac{S}{l}, \tag{B.1}$$

où S est la surface des électrodes, ϵ la permitivité diélectrique de la lame d'air d'épaisseur l. Compte tenu que S et ϵ sont constants, C est inversement proportionnel à l. Pour que le système soit opérationnel, il faut introduire une charge Q qui est fonction de la capacité

et de la tension V aux bornes des électrodes. La relation entre ces deux grandeurs est

$$Q = CV. \tag{B.2}$$

Les relations (B.1) et (B.2) permettent d'établir la relation

$$\Delta V = \frac{Q}{\Delta C} = \frac{Q\Delta l}{\epsilon S}, \tag{B.3}$$

qui montre que la variation de l'épaisseur de la lame d'air correspond aux variations de tension aux bornes du condensateur. Afin d'éviter l'action de la pression statique à l'intérieur de la cavité formée par les deux électrodes, un orifice entre cette cavité et l'environnement doit être ouvert. Dans les premiers prototypes un orifice unique a été implémenté, ce qui réduisait la résistance acoustique de la cavité. Certains auteurs ont alors optimisé les microphones avec des réseaux d'orifices gravés sur l'électrode fixe [Goto et al., 2007]. Le ratio de perforation doit être obtenu en faisant le compromis électromécanique, car il faut sacrifier la surface de l'électrode au profit de la souplesse de la membrane [Tajima et al., 2005] jouant ainsi directement sur la sensibilité du transducteur. La bande passante des microphones ainsi fabriqués a été augmentée jusqu'à 20 kHz.

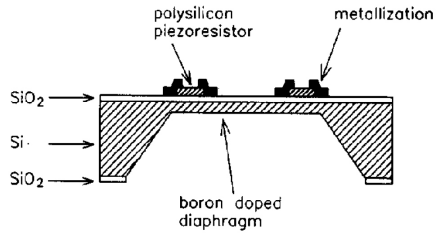
Initialement la membrane et la contre-électrode étaient fabriquées dans deux couches de silice indépendantes qui étaient ensuite collées. Cette technique présentait l'inconvénient d'avoir à souder les deux parties à des très hautes températures ce qui pouvait modifier les caractéristiques électromécaniques de la membrane et des électrodes [Scheeper et al., 2003]. La mise en œuvre de microphones réalisés sur une seule couche a été rendue possible grâce à la technique de la couche perdue [Tajima et al., 2005]. Avec cette technique il est possible d'éliminer une couche interne en plongeant le dispositif dans un dissolvant spécifique à la couche à éliminer. Cette technique permet de reproduire en masse des dispositifs présentant des performances techniques très proches.

B.3.2 Microphones piézoélectriques

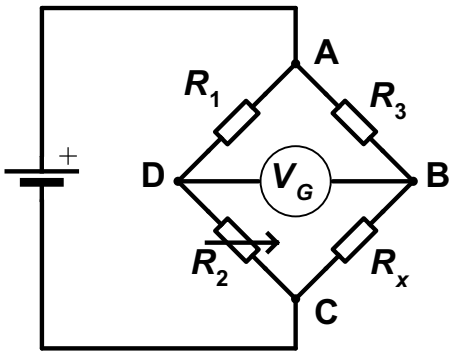
Les microphones piézoélectriques utilisent le changement des propriétés électriques des matériaux pour évaluer le déplacement de la membrane.

Ce type de microphones est divisé en deux classes :

- microphones piézorésistifs.



(a) Schéma de conception en coupe trasversale extrait de [Scheeper et al., 2003]



(b) Pont de Wheatstone

Figure B.2 : Exemple de réalisation d'un microphone piézorésistif.

- microphones piézoélectriques.

B.3.2.a Les microphones piézorésistifs

Pour ces microphones, la membrane mobile est munie de quatre jauges de déformation dans un matériau piézorésistif. Les jauges à résistivité variable sont branchées en pont de Wheatstone. Deux des jauges sont placées aux bords du diaphragme et deux au milieu de la membrane afin d'obtenir des résistivités équivalentes de signe opposé et ainsi pouvoir mesurer la déformation. L'un des premiers microphones réalisés utilisant cette technologie a été présenté par R. Schellin et G. Hess en 1992 comportant une membrane carrée de 0,5 μm d'épaisseur et 1 mm^2 de surface. Ce microphone a une sensibilité très faible de 25 μV/Pa et une réponse en fréquence relativement plate entre 100 Hz et 1 KHz.

B.3.2.b Les microphones piézoélectriques

Pour ces microphones, la membrane mobile est couverte d'un matériau piézoélectrique. Les contraintes mécaniques exercées sur le matériau engendrent des différences du potentiel électrique transmis par des électrodes situées de part et d'autre de la couche piézoélectrique.

Le premier des microphones *Micro-electro-mechanical Systems* (MEMS) mis en œuvre utilisant cette technologie, a été présenté par Royer [Royer et al., 1983] en 1983. L'ainé des microphones en silicium comportait un diaphragme circulaire de 30 μm d'épaisseur et de 3 mm de diamètre. Il a fourni une sensibilité assez faible, 25 μV/Pa et une réponse en fréquence entre 0,1 Hz et 10 KHz. Le rapport de signal sur bruit est seulement de 0,2 dB.

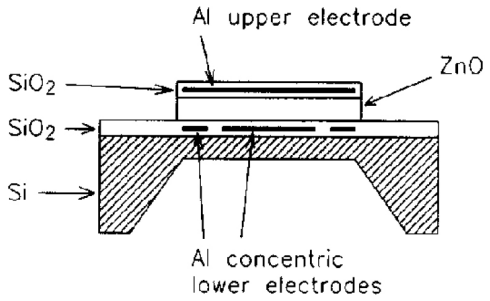


Figure B.3 : Schéma d'un microphone piézoélectrique extrait de [Scheeper et al., 2003].

Ces dispositifs sont assez bruyants et peu sensibles car la sensibilité dépend directement de la taille de la couche piézoélectrique et par conséquent de la taille de la membrane.

B.3.3 Microphones optiques

Le photophone a été inventé par G. Bell en 1881. Cette technologie présente l'avantage de pouvoir mesurer le déplacement de la membrane sans aucune intervention mécanique ou électrique, permettant une augmentation de la dynamique. Le principe de fonctionnement de ces dispositifs consiste à envoyer un signal lumineux monochrome d'intensité I et de longueur d'onde l à travers un réseau de réflecteurs. La lumière qui passe à travers les orifices du réseau est réfléchie sur la membrane mobile se trouvant à une distance d du réseau comme l'illustre la figure B.4.

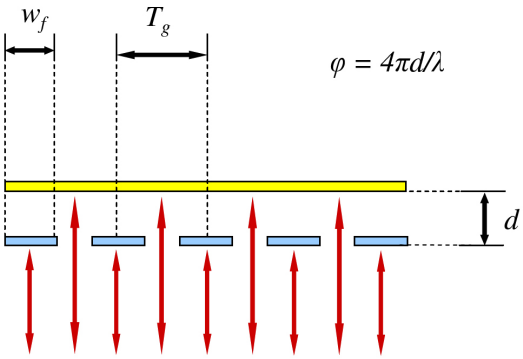


Figure B.4 : Principe d'interférométrie extrait de [Jeelani, 2009].

La lumière résultante est la superposition de la lumière réfléchie et diffractée par le système. L'intensité résultante s'exprime au premier et au deuxième ordre de la façon

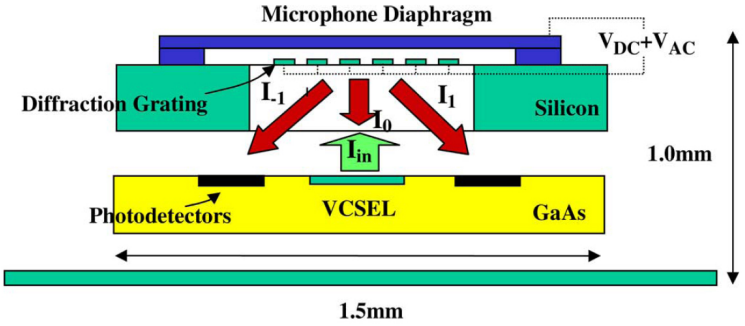


Figure B.5 : Schéma de conception de l'accéléromètre de Hall extrait de [Hall et al., 2008a].

suivante :

$$I_1 = I_{in} \cos^2\left(\frac{2\pi d}{\lambda}\right), \tag{B.4a}$$

et

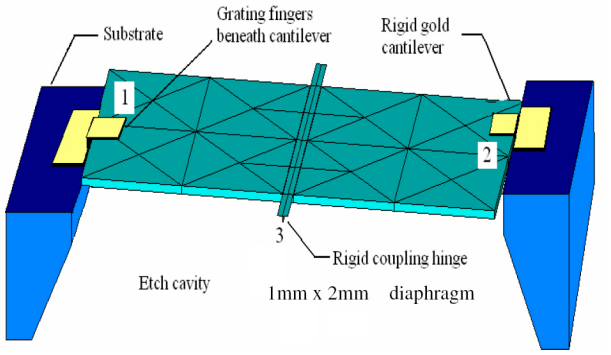
$$I_2 = \frac{4}{\pi^2} I_{in} \sin^2\left(\frac{2\pi d}{\lambda}\right). \tag{B.4b}$$

Elle est généralement captée par des photodiodes qui mesurent l'intensité lumineuse à un point donné. On peut alors déterminer le déplacement de la membrane en fonction de la variation de l'intensité. La sensibilité électrique et le domaine fréquentiel d'utilisation dépendent directement de la longueur d'onde de la lumière utilisée.

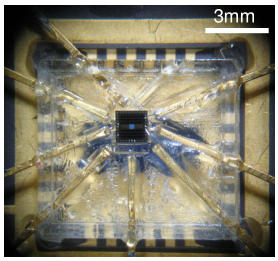
Initialement cette technologie a été implémentée sur des accéléromètres MEMS par Hall [Hall et al., 2008b] en 2008 et par la suite, le même auteur l'a utilisée dans la fabrication d'un microphone présentant une bande passante relativement plate entre 0 et 20 kHz [Hall et al., 2008a] selon le modèle illustré en figure B.5.

Jeelani [Jeelani, 2009] inspiré des travaux de Hall a voulu réaliser un microphone comportant une membrane rotative utilisant le mécanisme utilisé par le système auditif de la mouche *Ormia Ocharcea* (figure B.6). Ces expériences ont abouti à un système résonant à 1 kHz car le couplage acoustique-mécanique a été négligé dans la conception de l'appareil.

Le bruit dégagé par ce type de microphones est encore très élevé à cause des températures produites par les systèmes laser.



(a) Schéma de conception



(b) Microphone monté sur une puce éléctronique

Figure B.6 : Microphone *Biomimetic* de Jeelani [Jeelani, 2009].

B.3.4 Microphones FET (transistor à effet de champ)

Un transistor à effet de champ est un transistor à résistivité variable qui autorise le flux d'électrons en fonction du champ électrique environnant. Cette technologie a été utilisée pour la première fois dans le développement des microphones en silicium en 1991 par Künel [Hall et al., 2008b]. Pour obtenir la variation du champ la membrane a été métallisée. Le dispositif obtenu comportait une sensibilité de 0,2 mV/Pa et une réponse en fréquence relativement plate entre 100 Hz et 30 KHz. Ce dispositif a été amélioré en 1992 jusqu'à l'obtention d'une sensibilité de 5 mV/Pa.

Malgré la sensibilité obtenue, ces microphones sont peu répandus à cause de leur non-stabilité dans le temps et de leur sensibilité au bruit.

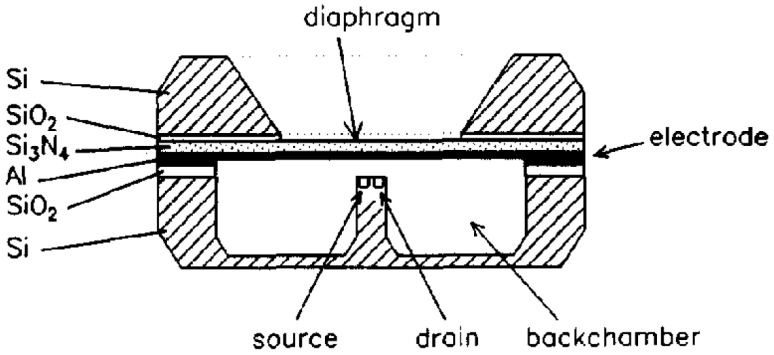


Figure B.7 : Schéma de conception d'un microphone FET extraite de [Scheeper et al., 2003].

B.4 Conclusion

Le développement des microphones de silicium a fait ses preuves dans la fabrication de microphones de taille conventionnelle [Scheeper et al., 2003] ainsi que dans la réalisation des microphones miniature. Initialement, les microphones ainsi réalisés présentaient une sensibilité très élevée aux perturbations électromagnétiques et une faible sensibilité. Les limitations des différentes technologies ont orienté les chercheurs sur l'optimisation des microphones capacitifs car cette technologie, contrairement aux autres, permet plusieurs degrés de liberté tant du point de vue mécanique qu'électronique. Afin de réduire considérablement les bruits occasionnés par les interférences électromagnétiques, les amplificateurs ainsi que les convertisseurs analogiques-numériques ont été intégrés dans une même puce. Du point de vue électro-acoustique, les microphones ainsi obtenus semblent fonctionner correctement. Néanmoins aucun test d'écoute des signaux enregistrés n'a été effectué. Ces microphones étant voués à intégrer les domaines de l'audio et des télécommunications, des tests subjectifs devront être entrepris pour évaluer leur qualité.

C Espace vectoriel L^2 et harmoniques sphériques

En coordonnées sphériques l'espace euclidien est repéré en azimut $\theta \in [0, 2\pi]$, en élévation $\phi \in [-\pi/2, \pi/2]$ et en fonction de la distance $r \in [0, \infty]$ par rapport à l'origine du repère.

Les relations qui lient coordonnées cartésiennes et sphériques sont

$$\begin{cases} x = r\cos\theta\cos\phi \\ y = r\sin\theta\cos\phi \\ z = r\ \sin\phi \end{cases} . \tag{C.1}$$

C.1 Produit Scalaire

Nous appelons L^2 l'espace vectoriel défini sur la sphère de rayon unitaire. Dans cet espace vectoriel nos pouvons définir le produit scalaire de deux vecteurs f et g par,

$$\langle f, g\rangle_S = \frac{1}{4\pi}\iint_{\theta,\phi} f(\theta,\phi)g(\theta,\phi)\cos\theta\,\mathrm{d}\phi\,\mathrm{d}\theta, \tag{C.2}$$

où

$$\frac{1}{4\pi}\cos\theta\,\mathrm{d}\phi\,\mathrm{d}\theta \tag{C.3}$$

représente une surface élémentaire de la sphère.

Le produit scalaire présente les propriétés suivantes

$$\begin{aligned} \langle f, f\rangle_S &\geq 0 \quad \forall \quad f \neq 0, \\ \langle f, f\rangle_S &= 0 \quad \text{ssi} \quad f = 0, \\ \langle f_1 + f_2, g\rangle_S &= \langle f_1, g\rangle_S \langle f_2, g\rangle_S, \\ \langle \alpha f, g\rangle_S &= \alpha \langle f, g\rangle_S, \\ \langle f, g\rangle_S &= \langle g, f\rangle_S. \end{aligned} \tag{C.4}$$

La norme d'un vecteur f est définie par,

$$\|f\| = \sqrt{\langle f, f\rangle_S} \tag{C.5}$$

et son énergie par,

$$\|f\|^2 = \langle f, f\rangle_S. \tag{C.6}$$

Dans L^2, deux éléments sont orthogonaux si,

$$\langle f, g\rangle_S = 0. \tag{C.7}$$

C.2 Base orthonormée

Si G est un ensemble de vecteurs appartenant à L^2, on dit que f est orthogonal à G si f est orthogonal à tous les éléments $g \in G$.

Si F est un ensemble de vecteurs appartenant à L^2, on dit que F est orthogonal à G si tout élément $f \in F$ est orthogonal à tous les éléments $g \in G$.

Une base orthonormée de L^2 est un ensemble de vecteurs orthogonaux deux à deux et de norme unitaire. L'ensemble de vecteurs $\mathcal{B} = (e_1, e_2 \cdots e_n)$ est une base orthonormée si et seulement si,

$$\begin{aligned} &\langle e_p, e_q \rangle_S = 1 \quad \text{ssi} \quad p = q \\ &\langle e_p, e_q \rangle_S = 0 \quad \forall \quad p \neq q \\ &\langle e_p, f \rangle_S = 0 \quad \text{ssi} \quad f = 0 \end{aligned} \tag{C.8}$$

avec $p, q \in \mathbb{N}$

C.3 Espace vectoriel

S'il existe une base orthonormée $\mathcal{B} = (e_0, e_1, \cdots, e_m \cdots)$, il est possible de définir un espace vectoriel $\mathbf{H}_M$ des fonctions h de la forme $h = \alpha_0 e_0 + \alpha_1 e_1 + \cdots + \alpha_M e_M$.

Il est possible d'exprimer tout vecteur $f \in L^2$ dans l'espace vectoriel $\mathbf{H}_M$ grâce à sa projection orthogonale :

$$f_M = \langle f, e_0 \rangle e_0 + \langle f, e_1 \rangle e_1 + \cdots \langle f, e_M \rangle e_M. \tag{C.9}$$

La norme $\|f_M - f\|$ tend vers 0 lorsque M tend vers l'infini. C'est-à-dire que f converge dans $L^2(S)$ lorsque N tend vers l'infini. L'erreur qui en résulte peut être rendue aussi faible que nécessaire en choisissant N suffisamment grand. On dit alors que $\sum \langle f, e_m \rangle e_m$ est le développement en série de f dans $L^2(S)$ par rapport à la base e_m, avec $m \in \mathbb{N}$.

C.4 Erreur d'approximation

La précision de l'approximation f_m est considérée par [Moreau, 2006] comme la différence d'énergie entre les fonctions f et f_M avec "l'erreur quadratique moyenne normalisée" $\overline{e_m}$. Cette dernière est définie suivant

$$\overline{e_M} = \frac{\|f - f_M\|_S^2}{\|f\|_S^2} = \frac{\int_{\theta=0}^{2\pi} \int_{\phi=-\pi/2}^{\pi/2} |f(\theta,\phi) - f_M(\theta,\phi)|^2 \cos\theta \, \mathrm{d}\phi \, \mathrm{d}\theta}{\int_{\theta=0}^{2\pi} \int_{\phi=-\pi/2}^{\pi/2} |f(\theta,\phi)|^2 \cos\theta \, \mathrm{d}\phi \, \mathrm{d}\theta}. \tag{C.10}$$

On obtient ainsi une valeur adimensionnelle normalisée entre 0 et 1.

Cette erreur détermine l'erreur moyenne commise sur l'ensemble de la sphère c'est pour cela que Moreau propose également le calcul de la variance associée.

$$\sigma_m^2 = \|e_M - \overline{e_M}\|_S^2 = \frac{1}{4\pi} \int_{\theta=0}^{2\pi} \int_{\phi=-\pi/2}^{\pi/2} |e_M(\theta,\phi) - \overline{e_M}|^2 \cos\theta \, \mathrm{d}\phi \, \mathrm{d}\theta. \tag{C.11}$$

D'autres auteurs proposent l'étude des valeurs maximales autour de la sphère comme alternative à la variance [Bamford, 1995], [Daniel, 2001] et [Poletti, 2000].

D DirAC

Cette méthode dont l'acronyme a pour signification "codage audio directionnel" *Directional Audio Coding* (DirAC) a été introduite par Pulkki [Pulkki, 2006]. Elle considère qu'à chaque instant le système auditif peut faire le distinguo entre une source sonore localisée et un champ sans direction déterminée (ou champ diffus).

Dans une première phase d'analyse, la méthode effectue la localisation des sources sonores sur des trames spectro-temporelles en effectuant une décomposition fréquentielle sur des bandes critiques ou *Equivalent rectangular bandwidth* (ERB). Le champ diffus est également estimé dans ce même découpage.

Cette discrimination permet de restituer le champ diffus sur l'ensemble des directions dans la phase de synthèse, au lieu de lui attribuer une direction privilégiée. De cette façon, la méthode prend en compte l'éventualité

- de la présence d'une source élargie dans l'espace,
- de plusieurs sources actives au même moment,
- d'un champs diffus créé par les réflexions dans une pièce.

L'analyse directionnelle est effectuée grâce aux signaux issus du format-B de l'ambisonique à l'ordre 1, à savoir (I.4.3.c) W (pression acoustique) et X, Y et Z, (composantes du vecteur de vitesse particulaire).

L'intensité acoustique est par définition, l'énergie qui traverse une surface pendant un instant donné [Bruneau, 1998]. Cette grandeur est le produit de la pression acoustique p et de la vitesse particulaire u, exprimé par

$$\vec{\Pi}(t) = p(t) \times u(\vec{t}) \,. \tag{D.1}$$

Ce vecteur complexe se décompose en deux parties, l'intensité acoustique réactive $\vec{J}$ et active $\vec{I}$,

$$\vec{\Pi}(t) = \vec{I} + i\vec{J}\,. \tag{D.2}$$

L'intensité acoustique réactive traduit les échanges locaux d'énergie (non propagatifs) et l'intensité acoustique active est une description vectorielle du transfert d'énergie acoustique. Cette grandeur est la moyenne temporelle de l'intensité instantanée sur une période dont la moyenne fluctuante est nulle. Ce vecteur s'exprime comme

$$\vec{I} = \frac{1}{2} Re(p^* \vec{u})\,, \tag{D.3}$$

où $*$ représente le complexe conjugué.

Cette expression devient dans le domaine fréquentiel et en coordonnées cartésiennes

$$\vec{I}(\omega) = \begin{bmatrix} I_x(\omega) \\ I_y(\omega) \\ I_z(\omega) \end{bmatrix} = \frac{1}{2} Re \left\{ P^*(\omega) \begin{bmatrix} V_x(\omega) \\ V_y(\omega) \\ V_z(\omega) \end{bmatrix} \right\}\,. \tag{D.4}$$

Dans le cas d'une onde plane, la pression acoustique peut être exprimée sous la forme

$$P(\omega) = P_o(\omega) e^{-i\vec{k}r}, \tag{D.5}$$

avec $\vec{k}$ le vecteur d'onde.

L'équation d'Euler exprime la vitesse acoustique comme une fonction de la pression acoustique grâce à l'impédance du milieu Z, de la forme

$$\vec{V}(\omega) = \frac{P(\omega)}{Z} \frac{\vec{k}}{\|\vec{k}\|}\,, \tag{D.6}$$

où $Z = \rho c$ avec ρ la masse volumique du milieu, c la célérité de l'air et $\|\quad\|$ la norme du

vecteur.

Les composantes du vecteur intensité deviennent alors

$$I_\mu(\omega) = \frac{|P_o(\omega)|^2}{2\rho c} \frac{k_\mu}{\|\vec{k}\|}, \tag{D.7}$$

où μ représente la composante géométrique x, y ou z.

Comme le vecteur d'onde $\vec{k}$ est colinéaire à la propagation de l'onde, ce dernier peut être utilisé pour estimer la direction de la source.

Le vecteur vitesse est extrait des signaux X, Y et Z de l'ambisonique grâce à la relation [Gerzon, 1975],

$$\vec{V}(\omega) = \frac{1}{\rho c} \begin{bmatrix} X(\omega) \\ Y(\omega) \\ Z(\omega) \end{bmatrix} . \tag{D.8}$$

Comme le signal W représente la pression acoustique, l'équation D.4 devient alors

$$\vec{I}(\omega) = \begin{bmatrix} I_x(\omega) \\ I_y(\omega) \\ I_z(\omega) \end{bmatrix} = \frac{1}{2\rho c} Re \left\{ W^*(\omega) \begin{bmatrix} X(\omega) \\ Y(\omega) \\ Z(\omega) \end{bmatrix} \right\} . \tag{D.9}$$

En fonction de la convention utilisée, il est nécessaire d'introduire dans l'équation D.9 un coefficient multiplicateur de $\sqrt{2}$ ou $1/\sqrt{2}$.

L'intensité acoustique est colinéaire à la direction de propagation de l'onde acoustique. Connaissant la direction du vecteur intensité on en déduit donc également la direction de la source. Cette dernière peut être calculée en azimut à partir de l'expression

$$\theta(\omega) = \tan^{-1} \left[\frac{I_y(\omega)}{I_x(\omega)} \right] , \tag{D.10}$$

et en élévation suivant

$$\phi(\omega) = \tan^{-1}\left[\frac{I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}}\right] . \tag{D.11}$$

Chacune de ces équations donnent deux résultats possibles à π près. Une direction unique est déterminée en fonction des signes des composantes du vecteur intensité.

La méthode *Directional Audio Coding* (DirAC) définit la partie diffuse ψ du signal comme la quantité d'énergie acoustique E oscillant localement [Merimaa and Pulkki, 2004]. Elle est calculée par la formule

$$\psi(\omega) = 1 - \frac{\|\langle \vec{I}(t)\rangle\|}{c\langle E(t)\rangle} = 1 - \frac{2Z\|\langle p(t)u(t)\rangle\|}{\langle p^2(t)\rangle + Z^2\langle u^2(t)\rangle} , \tag{D.12}$$

où $\langle \quad \rangle$ représente la moyenne et où

$$E(t) = \frac{1}{2}\rho\left[\frac{p^2(t)}{Z^2} + u^2(t)\right] . \tag{D.13}$$

La relation D.12 peut s'écrire dans le domaine fréquentiel sous la forme

$$\psi(\omega) = 1 - \frac{2Z\|Re\{P^*(\omega)V(\omega)\}\|}{|P(\omega)|^2 + Z^2|V(\omega)|^2} . \tag{D.14}$$

Lors de la phase de synthèse, les signaux sont diffusés sur des haut-parleurs réels ou virtuels placés aux directions obtenues lors de la première phase. Dans le cas où les haut-parleurs ne sont pas disponibles aux positions exactes, les signaux sont restitués en utilisant des alternatives de panoramique d'intensité comme le *Vector Based Amplitude Panning* (VBAP) [Pulkki, 2002] [Pulkki, 2001].

E Publications

E.1 Perceptual assesment of binaural decoding of first-order ambisonics [Palacino et al., 2012]

ACOUSTICS 2012

Perceptual assessment of binaural decoding of first-order ambisonics

J. Palacino, R. Nicol, M. Emerit and L. Gros

France Télécom - Orange Labs, FT/OLNC/RD/TECH/OPERA/TPS, 2 Av. Pierre Marzin, 22300 Lannion, France
julian.palacino@orange.com

The first-order Ambisonics microphone (e.g. Soundfield®) is a both compact and efficient set-up for spatial audio recording with the benefit of a full 3D spatialization. Another advantage is that the signals delivered by this microphone (i.e. B-Format) can be rendered over headphones by applying appropriate processing, while ensuring that the 3D spatial information is preserved. With the growing use of personal devices, it should be considered that most audio content is listened to over headphones. Thus first order Ambisonics recording provides an attractive solution to pick-up 3D audio content compatible with headphone reproduction. "Binaural decoding" refers to the processing to adapt B-Format for headphone rendering (i.e. "binaural format"). One solution is based on binaural synthesis of virtual loudspeakers. One promising way to improve the decoding is active processing which takes information from a pre-analysis of the sound scene, particularly in terms of spatial information. This paper will compare various binaural decoders. Starting from a listening test which assesses existing solutions and which shows that the perceived quality may strongly vary from one decoder to another, the processing is analyzed step by step. The performances are measured by a set of objective criteria derived from localization cues.

1 Introduction

3D audio recording provides immersive rendering of sound scene. Today Ambisonics (and its generalization Higher Order Ambisonics or HOA) proposes promising tools for spatial sound recording with the advantage of both full 3D spatialization and compact microphone set-up. On the contrary, 3D sound rendering remains a tricky issue, mainly in terms of equipment requirement. An attractive solution is therefore binaural processing of Ambisonics recordings, which means that the Ambisonics multichannel stream is adapted to headphone listening. In the following, this processing will be referred to as "binaural decoding" [1].

The objective of this paper is to assess the quality of binaural decoding. Various decoders are available today and the first step is a benchmark test to assess their performances, focusing on two strategies of decoding and, in addition, comparing them to other spatialization technologies. This test is presented in Section 2. Then the binaural decoding is analyzed step by step in Section 3, in order to identify where potential improvement may be found. Section 4 concludes the study by assessing the reconstruction of the signals delivered to the listener's ears for different options of processing.

2 Preliminary listening test

2.1 Objective

Our concern here is to provide spatial sound for headphone listening. Among the tools to record spatial sound, dummy-head is the most straightforward since binaural spatialization is precisely dedicated to headphone rendering. Stereophonic recording is another reference, as a conventional practice of sound engineer to record sound scene. Ambisonics proposes an attractive alternative. It should be highlighted that, in comparison to stereo, Ambisonics provides full 3D spatialization. However, it requires pos-processing, namely binaural decoding, to adapt the Ambisonics signals to headphone listening.

Thus, in a preliminary experiment, a listening test is performed in order to compare the perceived quality of various recordings of a spatial sound scene for the context of headphone listening. Three recording set-ups are considered: a dummy-head (KU100 Neumann acoustic head), an AB Stereo pair (i.e. a pair of 103 V 4003 DPA omni-directional microphones separated by 0.30 m) and a Soundfield® microphone. The test is based on excerpts taken from a live recording of the opera "Die Entführung aus dem Serail" of Mozart at the opera hall of the city of Rennes [2]. All the recording set-ups were placed above one seat and approximately at the potential location of the spectator's head, which allows the listener to be surrounded by the audience as he would be if he was really present in the hall. Two successive post-processing are applied to the output signals of the Soundfield® microphone: first conventional Soundfield® decoding to get the B-format signals [3], and second binaural decoding to adapt to headphone rendering. Two types of binaural decoders (which will be referred to as "SF dec1" and "SF dec2" in the following) are considered in our experiment, to contrast a "basic" decoder (SF dec1), i.e. mainly based only on the emulation of virtual loudspeakers, with an "active" decoder (SF dec2) in which the decoding is enhanced by sound scene analysis. On the contrary, the binaural recording obtained from the dummy-head is only equalized and the stereo signals are left untouched.

2.2 Experimental set-up

As a result, the listening test compares 4 types of sound spatialization, namely: dummy-head ("KU100"), stereo pair ("Stereo"), and 2 binaural decodings of the SoundField® signals ("SF dec1", "SF dec2"). The objective is to assess the overall quality (including both the audio and the spatial aspects) of the rendering of the sound scene over headphones. The experimental paradigm is based on a modified version of a MUSHRA test [4]. Since it is difficult to choose a priori one technology as a reference, no reference is proposed. Only one low anchor is added. This anchor includes both timbre and spatial degradations. Thus, for one trial, the subject is asked to judge a set of 5 pairs of signals ("KU100", "Stereo", "SF dec1", "SF dec2", "Anchor"). The assessment uses a multi criteria grid composed of 3 items: "quality", "space" and "timbre", following the methodology proposed in [5], except that a common anchor is used for each criterion. The overall test is based on ten audio excerpts covering various contents taken from the opera recording.

Twelve subjects (5 experts and 7 naive listeners) took part into the experiment. The overall test lasted around 2 hours and was divided into 2 parts separated by a break. The test interface was developed in Matlab. The experiment was carried out in an acoustically isolated room. The audio signals were presented to the subjects over HFI-580 Ultrasone closed headphones through a Terratec Phase 26 sound card configured for 48 kHz sampling rate and 24 bits resolution.

2.3 Results

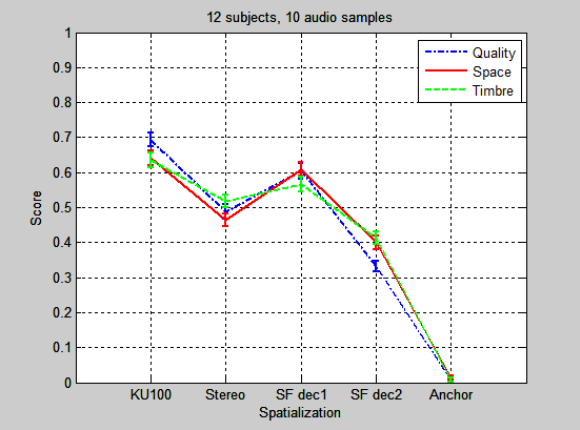


Figure 1: Mean score (and associated 95% confidence interval) in terms of quality, space and timbre.

For each sound spatialization, a total of 120 (10 audio excerpts x 12 subjects) scores were collected for the 3 criteria (quality, space, timbre). Figure 1 depicts the main score achieved by each technology. First, it should be noticed that the anchor of low quality was well identified. Second, in terms of global quality, the binaural recording is significantly preferred, followed by "SF dec 1" and the stereo pair. However, concerning space and timbre, the perception of the binaural recording and "SF dec 1" are very close. The stereo recording always stays in third position suggesting that this technology is not suited for headphone rendering. An ANOVA (ANalysis Of VAriance), considering 3 experimental factors (sound spatialization, audio excerpt, subject), confirms that the effect of sound spatialization is highly significant (p=0, F=233.89 for the "quality" criterion).

The same trend is observed for all subjects and all excerpts, except for the "applauses" excerpt. Indeed, for this latter, the binaural recording exhibits the worst score in terms of the "space" criterion. It may be due to the compression used to avoid the overload during the applauses. For all excerpts "SF dec 2" is judged significantly worse than "SF dec 1" in terms of "quality". It appears that binaural decoding of a first-order Ambisonic recording requires careful attention.

2.4 Conclusion

The results show that each system is clearly discriminated by the subjects and that the binaural rendering is significantly preferred. It is however striking that the Ambisonic recording is able to achieve a score close to the binaural sound, provided that a "proper" binaural decoding is applied. Indeed, it is also observed that the score of Ambisonics recording is highly dependant on the type of binaural decoding, which leads us to investigate the details of the processing in order to understand which element contributes to the perceived quality and where optimization can be expected.

3 Binaural decoding in question(s)

This part analyses step by step the overall processing from Ambisonics recording to headphone rendering.

3.1 HOA encoding

Ambisonics recording uses compact sensor arrays. The spatial encoding is based on the expansion of acoustical wave over spherical harmonics. Spherical harmonics Y_{mn}^{σ} define an Eigen base on the surface of a sphere of radius R defined by θ (azimuth) and ϕ $(elevation)$. Each element of this base is given by:

$$Y_{mn}^{\sigma} = \sqrt{(2m+1)\epsilon_n \frac{(m-n)!}{(m+n)!}} P_{mn}(sin\phi) \times \begin{cases} \cos n\theta \ si\ \sigma = 1 \\ \sin n\theta \ \ si\ \sigma = -1 \end{cases} \quad (1)$$

where

$$\begin{cases} m, n \in \mathcal{N}, n \leq m, \\ \sigma \in \{-1,1\}, \\ \epsilon_0 = 1, and\ \epsilon_n = 2\ if\ n > 0 \end{cases} \quad (2)$$

m is the harmonic order and P_{mn} are the associated Legendre functions defined in $x \in [-1,1]$ by:

$$P_{mn}(x) = (1-x^2)^{\frac{n}{2}} \frac{d^n}{dx^n} P_m(x) \quad (3)$$

Under the assumption that sound sources are outside of the sphere of radius R, the expression of the acoustic wave inside is done by the following expansion:

$$p(kr,\theta,\phi) = \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\phi) \quad (4)$$

where $j_m(kr)$ are spherical Bessel functions and B_{mn}^{σ} are obtained form the orthogonal projection of the acoustic pressure p to the corresponding spherical harmonic Y_{mn}^{σ}:

$$i^m j_m(kr) B_{mn}^{\sigma} = \langle p, Y_{mn}^{\sigma} \rangle \quad (5)$$

An approximation of pressure p can be done by truncating the expression (4) at the order $M \in \mathcal{N}$. This approximation gives K B_{mn}^{σ} coefficients defined by:

$$K = (M+1)^2 \quad (6)$$

The acoustic pressure field can be completely defined by the coefficients B_{mn}^{σ}. Those coefficients have been expressed here in the frequency domain. They can also be given in time domain using the inverse Fourier Transform.

$$b_{mn}^{\sigma}(t) = \int_{t=-\infty}^{\infty} B_{mn}^{\sigma}(\omega) e^{i\omega t}\, dt \quad (7)$$

In the case of a plane wave arriving from (θ_p, ϕ_p), which defines an angle γ with (θ,ϕ), the pressure is:

$$p(kr,\theta,\phi) = S\, e^{ikrcos\gamma} \quad (8)$$

its coefficients:

$$B_{mn}^{\sigma} = S(\omega) Y_{mn}^{\sigma}(\theta_p, \phi_p) \quad (9)$$

are the Ambisonics signals representing the whole acoustic field by the matrix relation:

$$\boldsymbol{b}_M = \boldsymbol{g}_M S(\omega) \quad (10)$$

where

$$\boldsymbol{b}_M = (B_{00}^1 \ldots B_{mn}^{\sigma} \ldots B_{MM-1}^{-1}) \quad (11)$$

and the gain array

$$\boldsymbol{g}_M = (Y_{00}^1 \ldots Y_{mn}^{\sigma} \ldots Y_{MM-1}^{-1}). \quad (12)$$

3.2 HOA decoding

The acoustic field can be reconstructed by a set L of loudspeakers located on the boundaries of a sphere. The coefficients B_{mn}^{σ} of the reconstructed soundfield are expressed as:

$$\boldsymbol{b} = \boldsymbol{C} \cdot \boldsymbol{s} \tag{13}$$

where

$$\boldsymbol{C} = \begin{bmatrix} Y_{00}^{1}(\theta_1,\phi_1) & \cdots & Y_{00}^{1}(\theta_l,\phi_l) & \ldots & Y_{00}^{1}(\theta_L,\phi_L) \\ \vdots & & \vdots & & \vdots \\ & & Y_{mn}^{\sigma}(\theta_l,\phi_l) & & \\ \vdots & & \vdots & & \vdots \\ Y_{M0}^{1}(\theta_1,\phi_1) & \ldots & Y_{M0}^{1}(\theta_l,\phi_l) & \ldots & Y_{M0}^{1}(\theta_L,\phi_L) \end{bmatrix} \tag{14}$$

and

$$\boldsymbol{s} = \begin{bmatrix} S_1 \\ \vdots \\ S_l \\ \vdots \\ S_L \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} B_{00}^{1} \\ \vdots \\ B_{mn}^{\sigma} \\ \vdots \\ B_{M0}^{1} \end{bmatrix} \tag{15}$$

S_l is the signal emitted by the loudspeaker l located at (θ_l, ϕ_l). The matrix $\boldsymbol{C}$ contains the spherical harmonics associated to the loudspeaker positions.

The objective is that the B_{mn}^{σ} coefficients of the reconstructed soundfield match those of the primary acoustic wave (Eq. (4)). An exact solution can be found when the number of loudspeakers L is higher than the order of truncation M. The loudspeaker signals are then derived from the B_{mn}^{σ} signals by applying a decoding matrix $\boldsymbol{D}$:

$$\boldsymbol{s} = \boldsymbol{D} \cdot \boldsymbol{b} \tag{16}$$

To find $\boldsymbol{D}$, which is in fact the inverse of the matrix $\boldsymbol{C}$, the Moore-Penrose pseudo-inverse can be used [6].

$$\boldsymbol{D} = \boldsymbol{C}^t \cdot (\boldsymbol{C} \cdot \boldsymbol{C}^t)^{-1} \tag{17}$$

If the loudspeaker array is uniformly distributed on the sphere, the relation (17) becomes [7]:

$$\boldsymbol{D} = \frac{1}{L}\boldsymbol{C}^t \tag{18}$$

because

$$\boldsymbol{C} \cdot \boldsymbol{C}^t = \frac{1}{L}\boldsymbol{I}_L \tag{19}$$

where $\boldsymbol{I}_L$ is the identity matrix of size $L \times L$.

3.3 Basic binaural decoding

Binaural synthesis uses a set of pair of binaural filters to create a virtual sound source for each position in space (r, θ, ϕ). Those filters are named Head Related Transfer Functions (HRTF) and can be obtained by measurement or modeling [8]. For Ambisonics decoding purposes, the reconstruction of acoustic field can be done by synthesizing virtual loudspeakers at the positions of available HRTFs. This method allows recreating an acoustic field at the entrance of the listener's ears.

The binaural signals $\boldsymbol{F_{bin}}$ of left and right channels are obtained as

$$\boldsymbol{F_{bin}} = \boldsymbol{H_{bin}} \cdot \boldsymbol{s} \tag{20}$$

$$\boldsymbol{F_{bin}}(\omega) = \begin{bmatrix} \boldsymbol{F_{bin,L}}(\omega) \\ \boldsymbol{F_{bin,R}}(\omega) \end{bmatrix} \tag{21}$$

$$\boldsymbol{H_{bin}}(\omega) = \begin{bmatrix} \boldsymbol{H_L}(\omega,\theta_1,\phi_1) & \ldots & \boldsymbol{H_L}(\omega,\theta_l,\phi_l) & \ldots & \boldsymbol{H_L}(\omega,\theta_L,\phi_L) \\ \boldsymbol{H_R}(\omega,\theta_1,\phi_1) & \ldots & \boldsymbol{H_R}(\omega,\theta_l,\phi_l) & \ldots & \boldsymbol{H_R}(\omega,\theta_L,\phi_L) \end{bmatrix} \tag{22}$$

where $\boldsymbol{H_{bin}}(\omega)$ is a matrix which defines the set of HRTFs measured for L directions. Substituting $\boldsymbol{s}$ for the loudspeaker signals in Eq.(20) leads to:

$$\boldsymbol{F_{bin}} = \boldsymbol{H_{bin}} \cdot \boldsymbol{D} \cdot \boldsymbol{b} \tag{23}$$

The binaural decoding matrix $\boldsymbol{E}$ is thus:

$$\boldsymbol{E} = \boldsymbol{H_{bin}} \cdot \boldsymbol{D} \tag{24}$$

which comes down to project the set of HRTFs on spherical harmonics [16]. Matrix $\boldsymbol{E}$ is $2 \times M$, which is a relatively small matrix for computation.

3.4 Pre-processing of HRTF

When implementing HRTF for binaural synthesis, some pre-processings are commonly used. It is intended to examine their potential impact on binaural decoding. First, modeling the HRTF by a minimum phase filter and a pure delay is considered. The delay is computed by the new method proposed by proposed by Nam [17] and validated by Nicol [18]. It consists in looking for the maximum of inter-correlation function between the HRIR and its minimum phase filter. Second, frequency smoothing is assessed. It is performed by critical band filter as described by Smith [19] and based on Hanning window.

In available databases (J.M. Pernaux, IRCAM[1], CIPIC[2], University of Maryland[3], Tohoku University[4] and Nagoya University[5]), HRTFs have been measured on the upper hemisphere of the sphere and in some cases also on part of the bottom hemisphere. The measurements are generally uniformly distributed over a single coordinate (azimuth or elevation) but not uniformly distributed over the whole sphere. However, for HOA decoding purposes, the projection of HRTF on spherical harmonics requires uniform sampling, in order to get an invertible decoding matrix in Eq.(19). Therefore HRTF interpolation is needed. The method based on Spherical Thin Plate Spline (STPS) derived from Wahba spherical spline [23] is chosen. Indeed Hartung et al [21] showed that this latter achieves the best performance.

4 Instrumental assessment of binaural decoding

Following the analysis of the processing, it is now intended to assess the performances of binaural decoding and to determine the effect of HRTF pre-processing over the resulting decoding. As a preliminary step, prior to a listening test, the assessment is based on a set of criteria, which are introduced in the next subsection.

4.1 Criteria

The assessment is focused on the rendering of spatial information. Therefore it is examined how the localization cues are reproduced in the signals delivered to the listener's ears. Sound localization uses mainly 2 kinds of cues: inter-aural cues (namely the Interaural Time Difference or ITD and the Inter-aural Level Difference or ILD, as described by the Lord Rayleigh's duplex theory [24]) and monaural cues [22]. ITD and ILD can be directly compared direction by direction. ILD is calculated using the method proposed by Larcher [16]:

$$ILD(\theta,\phi) = 10\log_{10}\frac{\int_{1.5kHz}^{10\ kHz}|H_L(\theta,\phi,f)|^2 df}{\int_{1.5kHz}^{10\ kHz}|H_R(\theta,\phi,f)|^2 df} \tag{25}$$

where H_L and H_R are the pressures at the left and right ear respectively. ITD is calculated using the method presented in subsection 3.4. Monaural cues rely essentially on spectral features. Therefore, the Inter-Subject Spectral Difference or

[1] http://recherche.ircam.fr/equipes/salles/listen/
[2] http://interface.cipic.ucdavis.edu/sound/hrtf.html
[3] http://www.isr.umd.edu/Labs/NSL/
[4] http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/
[5] http://www.sp.m.is.nagoya-u.ac.jp/HRTF/database.html

ISSD [25], which is derived from the variance of the difference between the original and reconstructed spectrum, is used to assess how spectral information (i.e. the frequency pattern involved in HRTFs) is correctly reproduced at the listener's ear:

$$\mathrm{ISSD}(\theta,\phi) = \left[\frac{1}{9\ kHz}\int_{4\ kHz}^{13\ kHz} 10\log_{10}\frac{\hat{H}(\theta,\phi,f)}{H(\theta,\phi,f)} - \Psi(\theta,\phi)df\right]^2 \quad (26)$$

$$\Psi(\theta,\phi) = \frac{1}{9\ kHz}\int_{4\ kHz}^{13\ kHz} 10\log_{10}\frac{\hat{H}(\theta,\phi,f)}{H(\theta,\phi,f)}\ df \quad (27)$$

For a set of HRTFs, a single value of ISSD can be calculated as the mean of ISSD of all considered directions. Middlebrooks points out the value of 6.18 dB as the optimum ISSD value [25].

4.2 Experimental protocol

For the evaluation of the different pre-processings, the private HRTF database *J.M. Pernaux* is used [20]. This database has a regular distribution on the upper part of sphere from an elevation of -56.25° and it contains 965 measured directions. The sampling frequency is 48 kHz and each HRTF is composed of 512 samples. All the assessment is performed over the subject labeled n° 1.

The various pre-processings are described in Table 1.

Table 1: List of applied pre-processings.

Description / Name	Minimum phase filter + ITD	Frequency smoothness	Interpolation
000			
100	X		
120	X	X	
130	X		X
123	X	X	X

Ambisonics encoding-decoding is applied over the entire set of HRTFs (965 original directions and 1026 interpolated directions). Spherical harmonic truncation used is 1, 4 and 30. The 1st and 4th orders corresponds to commercial Ambisonics microphones: 1st is Soundfield® [3] and 4th is Eigenmike®.The 30th order is calculated in order to get the best Ambisonics reconstruction as shown in Eq.(6) where optimum order M is chosen:

$$M = \sqrt{K} - 1 \quad (28)$$

K is the number of measured directions.

4.3 Experimental Results

Monaural cues (ISSD)

As shown in Figure 2, decreasing reconstruction order corrupts principally high-frequencies. Table 2 lets appear that Ambisonics reconstruction is quite precise at 30th order. In addition, processing "100" improves its quality in comparison to direct Ambisonics reconstruction for 30th and 4th order. But ISSD is degraded for smoothed HRTF and this occurs before Ambisonics processing. Nevertheless ISSD is lower for an Ambisonics reconstruction of any order of a smoothed HRTF. This property can be used to simplify HRTF spectrum for computational saving.

In the present case, interpolation over non-measured directions adds a negligible improvement of Ambisonics reconstruction. The interest of interpolation can be discarded taking into account that this pre-processing is computational expensive. However the current HRTF database is regularly distributed. Further test must be done over HRTFs sets that don't have this property.

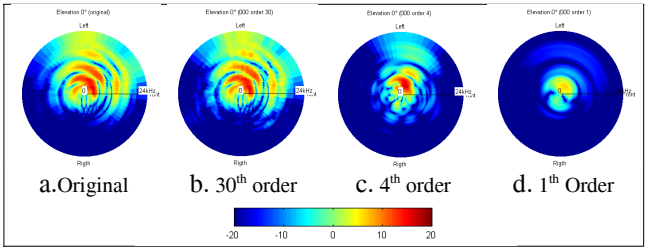


Figure 2: Horizontal cut at elevation 0° of magnitude spectrum of HRTF, Original (a.) and 1st ,4th, 30th order Ambisonics reconstruction without pre-processing

Table 2: ISSD values for different orders and pre-processing methods (in dB^2).

Pre-processing Name	Ambisonics order 1: ISSD (dB^2)	1: ISSD Degradation (dB^2)	4: ISSD (dB^2)	4: ISSD Degradation (dB^2)	30: ISSD (dB^2)	30: ISSD Degradation (dB^2)	Pre-processing only (Before Ambisonics decoding)
0	39.04	*39.04*	45.24	*45.24*	9.69	*9.69*	0
100	39.91	*39.38*	33.26	*32.73*	3.68	*3.15*	0.53
120	39.55	*30.82*	33.70	*24.97*	10.94	*2.21*	8.73
130	39.55	*39.02*	32.98	*32.45*	4.81	*4.28*	0.53
123	39.15	*30.42*	33.39	*24.66*	11.7	*2.97*	8.73

Binaural cues (ITD, ILD)

Table 3: ITD error mean and uncertainty values for 30th order encoding-decoding (in μs).

Name	000	100	120	130	123	mean
Mean ITD Error (μs)	18.8	14.9	14.6	15.0	14.8	15.6
ITD Error std. deviation (μs)	*2.5*	*3.6*	*3.9*	*3.7*	*3.9*	*3.5*

ITD varies commonly between 0 and 700 μs. As shown in Table 3, the ITD is well reconstructed at 30th order. Error is Gaussian distributed over a mean value varying around 15 μs with a standard deviation of 3 μs. For lower orders (1st and 4th) and for all the pre-processings, the ITD is completely lost and the resulting value oscillates around 0 s.

Like ISSD, ILD is really well reconstructed at 30th order. The achieved error is always less than 1 dB. At 4th order reconstruction of non pre-processed ILD values is Gaussian distributed over a mean value varying around 0 dB with a standard deviation of 3 dB and for some directions the maximum error reaches 8 dB. Anyway the spatial variation is coherent with natural ILD. All pre-processings increase the ILD error mainly for directions at the north hemisphere where ILD is over estimated. The mean of absolute value ITD error is then 7 dB and its standard deviation is 3 dB.

Generally 1st order reconstruction of non pre-processed HRTF gives good results in terms of ILD. Only some values at spots situated at (90°,-20°) and (270°,-20°) are over estimated. All pre-processings increase the number of maximum areas of ILD error. This happens because energy is focused principally over the Eigen vectors of 1st order spherical harmonics.

5 Conclusion

In the present paper, we studied the impact of HRTF database pre-processing for Ambisonics encoding-decoding purposes. Pre-processing considered are: modeling the HRTF by a minimum phase filter and a pure delay, frequency smoothing and HRTF interpolation over a regular distributed HRTF set over the whole sphere. HRTF reconstruction was assessed in terms of monaural cues using ISSD and in terms of interaural cues using ILD and ITD.

Modeling the HRTF by a minimum phase filter and a pure delay gives best results in terms of ISSD. Smoothness deteriorates ISSD before Ambisonics reconstruction but the reconstruction in any order of a smoothed HRTF is better than of non smoothed HRTF. Interpolation doesn't provide any improvement for the current HRTF database. Further studies must be done over less regular distributed HRTF sets.

ILD and ITD are well reconstructed over 30^{th} order and for all studied pre-processings. For lower orders, ILD is over estimated for some directions buts its general behavior remains. On the contrary, ITD is completely lost for 1^{st} and 4^{th} orders.

Future work will investigate the evolution of the different criteria as a function of Ambisonics order and the perceptual links by a listening test. Binaural active decoding is another issue to examine with the same criteria.

References

[1] S. MOREAU, « Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance », École doctorale de l'université du Maine, Le Mans, France, 2006.

[2] http://www.mozartsurecrans.com/ (consulted on September 2011)

[3] M.A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound." *50th AES International conference*, 1975.

[4] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunications Union, Radio-communication Assembly*, 2003

[5] S. Le Bagousse et al., "Sound Quality Evaluation based on Attributes - Application to Binaural Contents ",131th AES International conference. New York, USA, 2011

[6] Golub et al, "Matrix computations". 3th edition. JHU Press, 1996.

[7] D. B. Ward et al. "Reproduction of a plane-wave sound field using an array of loudspeakers ", *Speech and Audio Processing, IEEE Transactions on*, vol. 9, n°. 6, p. 697–707, 2001.

[8] R. Nicol, "Binaural technology". *Audio Engineering Society*, 2010.

[9] Kulkarni, A., et al. "Sensitivity of human subjects to head-related transfer-function phase spectra". J. *Acoust. Soc. Am*. 105 (1999): 2821.

[10] Mehrgardt (S.) et al, "Transformation characteristics of the external human ear", *J. Acoust. Soc. Am*., 61(6), 1977, p. 1567–1576.

[11] Glasberg (B. R.) et al, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, 47, 1990, p. 103–138.

[12] Asano (F.) et al, "Role of spectral cues in median plane localization", *J. Acoust. Soc. Am*., 88(1), 1990, p. 159–168.

[13] Kulkarni (A.) et al, "Variability in the characterization of the headphone transfer-function", *J. Acoust. Soc. Am*., 107(2), 2000, p. 1071–1074.

[14] Guillon, Pierre. « Individualisation des indices spectraux pour la synthèse binaurale ». *Ph.D, Université du Maine*, 2009.

[15] P. Minnaar, et al, "The interaural time difference in binaural synthesis", presented at the *AES 108th convention*, Paris, 2000.

[16] V. Larcher, « Techniques de spatialisation des sons pour la réalité virtuelle », *PhD, Paris VI*, Paris, 2001.

[17] J. Nam, et al, "A method for estimating interaural time difference for binaural synthesis", in *125th Audio Engineering Society Convention*, San Francisco, 2008, vol. 21.

[18] R. Nicol, « Représentation et perception des espaces auditifs virtuels », *HDR, Université du Maine*, Le Mans, France, 2010.

[19] J. O. Smith, "Techniques for Digital Filter Design and System Identification with Application to the Violin", *PhD thesis, Elec. Engineering Department., Stanford University* (CCRMA), June 1983.

[20] J.-M. PERNAUX, « Spatialisation du son par les techniques binaurales : application aux services de télécommunications », PhD, I.N.P.G, Grenoble, 2003.

[21] K. Hartung et B. Jonas, "Comparison of different methods for the interpolation of head-related transfer functions", in *AES 16th*.

[22] J. Blauert, "Spatial hearing: the psychophysics of human sound localization", 2^{e} éd. Cambridge: MIT Press, 1983.

[23] G. Wahba, "Spline Interpolation and smoothing on the sphere", *SIAM J. Sci. Stat. Comput.*, vol. 2, mars 1981.

[24] L. Rayleigh, "On our perception of sound direction", *philosophical Magazine*, 13, 1907, p. 214-232

[25] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency", *The Journal of the Acoustical Society of America*, vol. 106, p. 1480, 1999.

E.2 Full 3D sound pick-up with a small microphone array : Prototype outline and preliminary assessment [Palacino and Nicol, 2013a]

Full 3D sound pick-up with a small microphone array: Prototype outline and preliminary assessment

Julian Palacino[1], Rozenn Nicol[2]

[1] *Orange Labs, 22300 Lannion, France, Email: julian.palacino@orange.com*
[2] *Orange Labs, 22300 Lannion, France, Email: rozenn.nicol@orange.com*

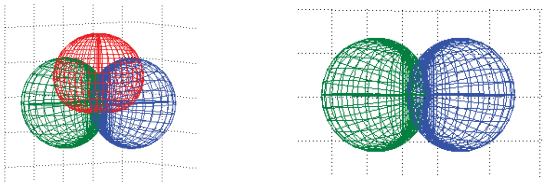
Introduction

For several decades spatial audio has been only used by movies, music composers and researchers in laboratories [1][2]. Because of the complexity of 3D audio technologies, people have always been away from these techniques. Dedicated devices such as microphone and loudspeaker arrays are expensive and cannot be used without some expertise of audio capturing and reproduction. Nowadays the main barrier dissuading of a consumer solution from capturing spatial audio is the big number of transducers needed to get an accurate 3D sound image [3][4][5]. In order to break down this barrier we propose a new 3D audio recording set-up which is composed of a three-microphone array able to get the full 3D audio information. A 2D version, consisting of a two-microphone array, is also available. Nowadays most of techniques of sound localization make the assumption that microphones are omnidirectionals and time delays are used as the main cue [6][7]. In order to reduce the number of transducers the sound localization method described in this paper is based on the transducer directivities and time delay is used as additional information allowing solving the angular ambiguity. The paper will describe first the microphone set-up and its associated algorithm. Secondly the performances of sound localization will be assessed.

Source localization using microphone directivity pattern

Microphone Array Layout

The microphone array is composed of 3 cardioid microphones (see Fig. 1): the first one pointing to the x axis (right), the second one to the opposite direction (left or -x) and the third one to the z axis (top).



a. 3D array *b. 2D array*

Fig. 1 Layout of the microphone device

Spatial Information Achieved from Microphone Directivity

The method operates in the time/frequency domain. It is assumed that only one sound source is present at each moment for a single frequency bin. Before applying FFT, time samples are weighted by a soft edge window to avoid oscillations in frequency domain. In terms of signal processing, the choice of the window type and its length is important. Particularly the slope and length of the time window will result in frequency spreading. To avoid instability of the source localization, results can be smoothed both frequency and time wise [8]. In the following, equations are presented for a fixed time-frequency bin.

The directivity of the n[th] cardioid microphone is represented by

$$\mathrm{M_n}(\alpha_\mathrm{n}) = \frac{1}{2}(1+\alpha_\mathrm{n}) \qquad (1)$$

where

$$\alpha_\mathrm{n} = \overrightarrow{d_s}\cdot\overrightarrow{d_{pn}} \qquad (2)$$

The vector $\overrightarrow{d_s}$ defines the source direction and the vector $\overrightarrow{d_{pn}}$ refers to the pointing direction of n[th] microphone.

In this case, the pointing direction $\overrightarrow{d_{pn}}$ can be expressed in the Cartesian basis $\mathcal{B}_c$ for each microphone by

$$\overrightarrow{d_{p1}} = \begin{vmatrix}1\\0\\0\end{vmatrix}_{\mathcal{B}_c}, \quad \overrightarrow{d_{p2}} = \begin{vmatrix}-1\\0\\0\end{vmatrix}_{\mathcal{B}_c}, \quad \overrightarrow{d_{p3}} = \begin{vmatrix}0\\0\\1\end{vmatrix}_{\mathcal{B}_c} \qquad (3)$$

Source location direction $\overrightarrow{d_s}$ can be expressed in the Spherical or Cartesian coordinate basis, respectively $\mathcal{B}_s$ or $\mathcal{B}_c$, by

$$\overrightarrow{d_s} = \begin{vmatrix}\theta\\\phi\\r\end{vmatrix}_{\mathcal{B}_s} = \begin{vmatrix}r\cos\phi\cos\theta\\r\cos\phi\sin\theta\\r\sin\phi\end{vmatrix}_{\mathcal{B}_c} \qquad (4)$$

where the Spherical coordinates are defined by radius r, azimuth angle θ, and elevation angle ϕ.

The directivity functions of the three microphones is given by [10]:

$$M_{card1}(\theta,\phi) = \tfrac{1}{2}(1+r\cos\phi\cos\theta), \quad a$$
$$M_{card2}(\theta,\phi) = \tfrac{1}{2}(1-r\cos\phi\cos\theta), \quad b \qquad (5)$$
$$M_{card3}(\theta,\phi) = \tfrac{1}{2}(1+r\sin\phi) \quad c$$

Since the direction is unchanged for any value of r, radius is fixed to $r = 1$ in following expressions.

Assuming that all microphones of the array are located at the basis origin (0,0,0) their output signals $s_{cardn}(\mathrm{t})$ are:

$$s_{cardn}(t,\theta,\phi) = M_{cardn}(\theta,\phi)s_0(t) \qquad (6)$$

where $s_0(t)$ is the acoustic pressure at this point and M_{cardn} the microphone directivity. This last acts as a gain factor which is function of the sound source location.

The signals $\mathrm{s_{cardn}(t,\theta,\phi)}$ allow to getting three data:

1) The monophonic signal or the sound pressure level using relations (5).a (5).b and (6)

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi) \quad (7)$$

2) The elevation angle of the sound source, by using equations (5).c and (6)

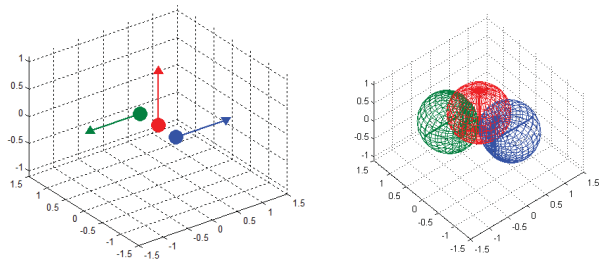
$$\phi = \sin^{-1}\left[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1\right] \quad (8)$$

3) The azimuth angle of the sound source, by using relations (5).a (5).b

$$\theta = cos^{-1}\left[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}\right] \quad (9)$$

Alternately it is possible to use other directivity pattern microphones to reconstruct cardioid directivity virtually [9]. The same method can be used on a 2D version restricted to two microphones corresponding to the horizontal plane (see Fig. 1a.). In this case an arbitrary elevation must be fixed. This introduces an error increasing with the angular mismatch from the chosen elevation and the real one.

In this first step, source location is calculated using exclusively the microphone directivity pattern. However the azimuth is estimated with a sign ambiguity (front-rear) due to the cosine of Equation (9). This ambiguity can be solved by moving microphones perpendicularly to their pointing direction, which will be illustrated in the next section (see Fig. 2.)



a *b*

Fig. 2 Layout of the non coincident microphone device, a) location and pointing directions, b) directivity patterns

Front – Rear Ambiguity Resolution Using Time Delay

If we consider now that the n[th] microphone is located at the position defined by the $\overrightarrow{E_n}$ vector:

$$\overrightarrow{E_n} = \begin{matrix} x_n \\ y_n \\ z_n \end{matrix}\Big|_{\mathcal{B}_c} \quad (10)$$

its output signal, s_{cardn}, becomes:

$$s_{cardn}(t,\theta,\phi) = M_{cardn}(\theta,\phi) \times s_0(t-\tau_n) \quad (11)$$

where τ_n is the delay induced by the distance between the n[th] microphone and the basis origin given by

$$\tau_n = \frac{1}{c}\overrightarrow{d_s}\cdot\overrightarrow{E_n} \quad (12)$$

and it is assumed that the sound source is in the far field.

In the frequency domain $s_{cardn}(t)$ becomes $S_{cardn}(\omega)$ (where $\omega = 2\pi f$ is the angular frequency with f the time frequency) by applying a Fourier Transform $FT[\ ]$

$$S_{cardn}(\omega,\theta,\phi) = FT[s_{cardn}(t,\theta,\phi)] = M_{cardn}(\theta,\phi)S_0(\omega)\, e^{-j\omega\tau_n}, \quad (13)$$

with

$$S_0(\omega) = |S_0(\omega)|\, e^{j\angle S_0(\omega)}. \quad (14)$$

Equation (13) becomes

$$S_{cardn}(\omega,\theta,\phi) = M_{cardn}(\theta,\phi)|S_0(\omega)|\, e^{j(\angle S_0(\omega) - \omega\tau_n)} \quad (15)$$

Consequently

$$\angle S_{cardn}(\omega,\theta,\phi) = \angle S_0(\omega) - \omega\tau_n \quad (16)$$

and

$$\angle S_{\mathrm{card1}} - \angle S_{\mathrm{card2}} = -\omega(\tau_1 - \tau_2). \quad (17)$$

The time delay $\tau_{12} = \tau_1 - \tau_2$ between the two microphones is expressed by:

$$\tau_{12} = -\frac{1}{\omega}(\angle S_{\mathrm{card1}} - \angle S_{\mathrm{card2}}). \quad (18)$$

Since this result is used to solve the ambiguity in relation (9) only its sign is needed. It is inserted in this latter as:

$$\theta = \frac{\tau_{12}}{|\tau_{12}|} cos^{-1}\left[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}\right] \quad (19)$$

As a time delay is introduced spatial aliasing is presented over wavelengths close to microphone spacing. Consequently the reproduction of $s_0(t)$ given in the relation (7) is not longer valid. In order to reduce this effect, it is necessary to use the energy of the signal. Consider the Fourier Transform function as an integral in the time domain over a temporal window. It is necessary to choose the window length as bigger as the traveling time between the farthest microphones. In this case al the terms s_{cardn} in the relations (7)(8)(9) must be replaced by their energy values $E(s_{cardn})$

So in the 2d array case, azimuth is determined by the power spectrum difference of the microphone outputs normalized by the power spectrum of the sound pressure level.

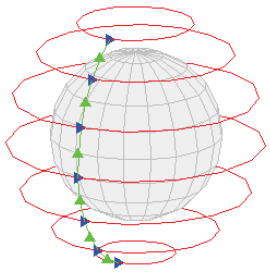


Fig. 3 Virtual sound source trajectory: azimuth (blue arrows) and elevation (green arrows)

Localization performances

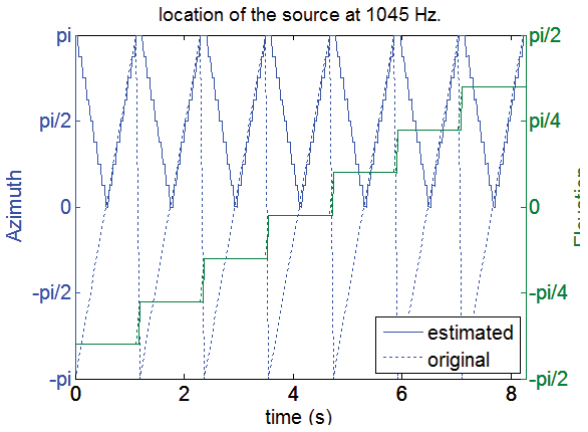


Fig. 4 Source Localization of a random noise at 1 kHz in azimuth (blue) and elevation (green) picked up with a coincident cardioid microphone array. Horizontal plane microphones are separated by 2cm.

A computer program synthesizes the signals which would have been recorded by the 3D array previously described. Various *stimuli* were used (music, random noise, noise band and harmonic tone). Figures presented in this section have been obtained using a random noise moving around and from bottom to the top as shown in Fig. 3.

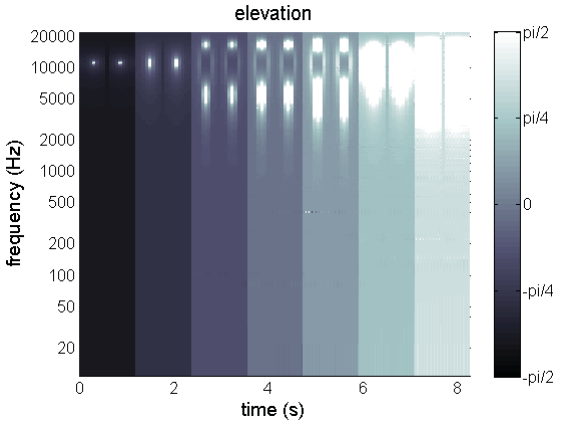


Fig. 5 Elevation localization of a random noise (trajectory illustrated in Fig. 3) picked up with a cardioid microphone array. Horizontal plane microphones are separated by 2cm. The estimated elevation is represented as the function of time and frequency (gray scale depicted on the right)

It is observed that when an ideal coincident array is used, the sound source is localized with front-rear ambiguity as depict in Fig. 5. Apart from the front rear ambiguity, location is well estimated for every frequency.

Since a time delay is introduced spatial aliasing degrades elevation localization over wavelengths close to the microphone distance when relations (7) and (8) are used. Localization accuracy is poor for elevation higher to $\pi/4$ because the reconstruction of the omnidirectional pressure S_0 is slightly affected by the microphone spacing. Since in high elevations the dynamic of the level differences of the third cardioid pattern (s_{card3}) is lower, the angular discrimination is poorer. As a result, the azimuth estimation is also degraded indeed the relation (9) shows how elevation influences the azimuth estimation.

If the energetic approach is used in combination with an appropriate choice of the integration window, spatial aliasing doesn't impact the performances of the elevation localization. The azimuth is also well detected as shown in Fig. 6. Only a few front - rear ambiguities remain. To avoid instability of the source localization, results can be smoothed frequency and time wise [8].

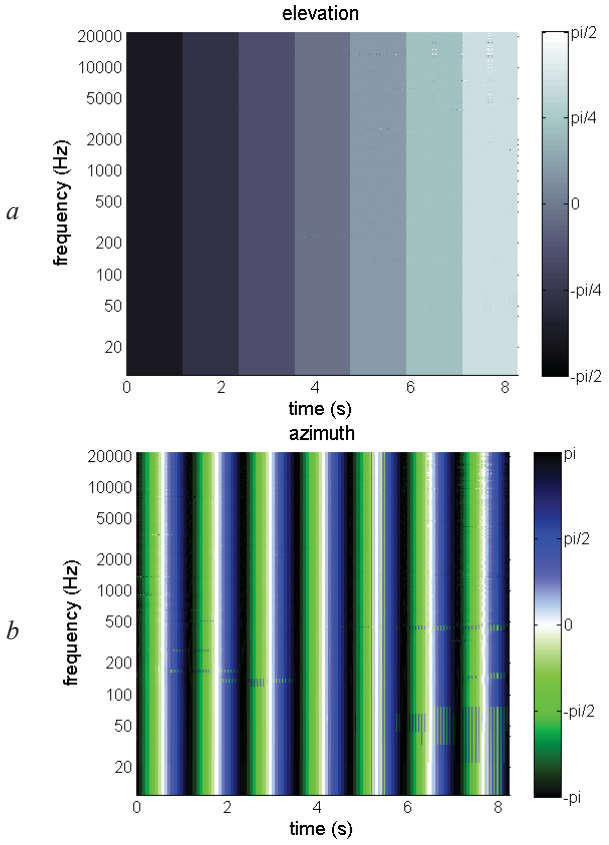
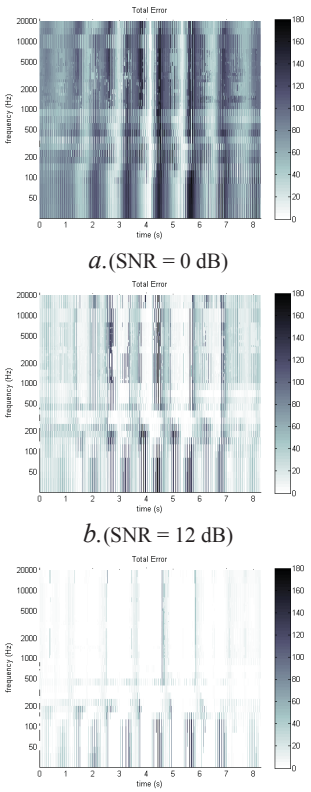


Fig. 6 Elevation (a) and azimuth (b) localization of a random noise picked up with a cardioid microphone array and based on energy estimation ($E(s_{cardn})$). Horizontal plane microphones are separated by 2cm.

Effects of a disturbing source

Fig. 7, illustrates the localization error as the angular distance between the real location and the estimated one in presence of a disturbing source located in front of the array. Localization performances are also affected when two or more sources are at different locations in the same frequency band. Low energy signals spatially close to high energy signals are then localized to the direction of the louder one. For a Signal to Noise Ratio (SNR) close to 0 dB, a single source is localized randomly between the two sources. When the SNR reaches 10 dB front rear instabilities happened for any position of the target source. The worst case is achieved when the target and the disturbing source are diametrically opposed. Azimuth of the target source is then detected with an error lower than 23°. For SNR is higher than 20 dB the source is well localized and front rear ambiguities remains only on low frequencies.



a.(SNR = 0 dB)

b.(SNR = 12 dB)

c.(SNR = 20 dB)

Fig. 7 Localization error of a random noise picked up with a cardioid microphone array in the presence of a disturbing source located at ($\boldsymbol{\theta = 0, \phi = 0}$). SNR is a.0 dB, b. 12 dB and c. 20 dB

Conclusion and further works

In order to provide 3D audio tools for consumer we propose a recording solution using a three-microphone array. The omnidirectional pressure magnitude is recomposed and compared with the directional microphone output. Sound source location is estimated in azimuth and elevation using microphone directivity. However the localization is front-rear ambiguous. This ambiguity is the solved by shifting the two horizontal microphones in order two introduce a time delay depending on the sound direction hemisphere.

In the case of non-coincident arrays, the omnidirectional pressure component is not properly reconstructed at wavelengths closer to microphone spacing. Indeed spatial aliasing alters the localization. This can be solved by focusing on the signal energy, allowing getting accurate localization over all frequency range, provided that the window size is adapted to the microphone size.

This study has been carried on in the case of ideal microphone arrays by computational simulations. The next step is to study the performances of real devices and how their deviation in terms of directivity, frequency response and self noise production affects localization accuracy in anechoic and real environments.

The 3D microphone array output can be seen as a kind of object based representation of the spatial sound scene [11][12]. Therefore using the information of source position, it is possible to render the sound scene over any type of spatial audio system such as stereo, 5.1, 7.1, 22.2, Higher Order Ambisonics [13] or Wave Field Synthesis [14].

References

[1]. J. Sunier, The story of stereo: 1881-. Gernsback Library, 1960.

[2]. F. Rumsey, Spatial Audio. Taylor & Francis, 2001.

[3]. Michel A. Gerzon et Peter G. Craven, « Coincident microphone simulation covering three dimensional space and yelding various directional outputs », U.S. Patent 4,042,77916-août-1977.

[4]. G.Theile, « Natural 5.1 music recording based on psychoacoustic principals », in Audio Engineering Society Conference: 19th International Conference: Surround Sound-Techniques, Technology, and Perception, 2001.

[5]. S. MOREAU, « Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance », ÉCOLE DOCTORALE DE L'UNIVERSITÉ DU MAINE, LE MANS, FRANCE, 2006.

[6]. R. Schmidt, « Multiple emitter location and signal parameter estimation », Antennas and Propagation, IEEE Transactions on, vol. 34, n° 3, p. 276–280, 1986.

[7]. E. Vincent, H. Sawada, P. Bofill, S. Makino, et J. Rosca, « First stereo audio source separation evaluation campaign: data, algorithms and results », Independent Component Analysis and Signal Separation, p. 552–559, 2007.

[8]. S. Berge et N. Barrett, « A new method for B-format to binaural transcoding », in AES 40th international conference, Tokio, Japan, 2010.

[9]. J. Palacino et R. Nicol, « Full 3D sound pick-up with a small microphone array », in ICA 2013, Montreal, 2013.

[10]. J. Jouhaneau, Notions élémentaires d'acoustique: Électroacoustique. Tec & Doc Lavoisier, 1999.

[11]. V. Pulkki, « Directional audio coding in spatial sound reproduction and stereo upmixing », in Proc. of the AES 28th Int. Conf, Pitea, Sweden, 2006.

[12]. M. Goodwin et J. M. Jot, « Spatial audio scene coding », in AES 125th Convention, San Francisco, CA, USA, 2008.

[13]. J. Daniel, « Evolving views on HOA: From technological to pragmatic concerns », Ambisonics Symposium 2009, June 25-27, Graz, 2009.

[14]. A.J. Berkhout, D. de Vries & P. Vogel, « Acoustic Control by Wave Field Synthesis », J. Acoust. Soc. Am., 1993, 93, pp. 2764-2778.

E.3 Spatial sound pick-up with a low number of microphones [Palacino and Nicol, 2013b]

Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 4aSP: Sensor Array Beamforming and Its Applications

4aSP2. Spatial sound pick-up with a low number of microphones

Julian D. Palacino* and Rozenn Nicol

***Corresponding author's address: SVQ//TPS, Orange Labs, 2 Av Pierre Marzin, Lannion, 22307, Britany, France, julian.palacino@orange.com**

For several decades spatial audio has been only used by movies, music composers and researchers in laboratories. Because of their complexity, people have always been away from 3D audio techniques. Dedicated devices such as microphone and loudspeaker arrays are expensive and cannot be used without some expertise of audio capturing and reproduction. Nowadays the main barrier preventing a consumer solution from capturing spatial audio is the big number of transducers needed to get an accurate 3D sound image. In order to break down this barrier we propose a new 3D audio recording set-up which is composed of a three-microphone array able to get the full 3D audio information. A 2D version, consisting of a two-microphone array, is also available. The sound localization is based on the transducer directivities and additional information to solve the angular ambiguity. This paper will describe firstly the microphone set-up and its associated algorithm. Secondly the performances of sound localization will be assessed.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4800844]
Received 22 Jan 2013; published 2 Jun 2013
Proceedings of Meetings on Acoustics, Vol. 19, 055078 (2013)

INTRODUCTION

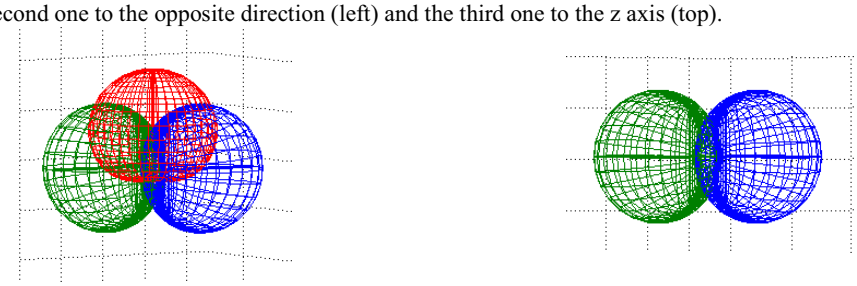
For several decades spatial audio has been only used by movies, music composers and researchers in laboratories [1]. Because of their complexity, people have always been away from 3D audio techniques. Dedicated devices such as microphone and loudspeaker arrays are expensive and cannot be used without some expertise of audio capturing and reproduction. Nowadays the main barrier preventing a consumer solution from capturing spatial audio is the big number of transducers needed to get an accurate 3D sound image [2]. In order to break down this barrier we propose a new 3D audio recording set-up which is composed of a three-microphone array able to get the full 3D audio information. A 2D version, consisting of a two-microphone array, is also available. The sound localization is based on the transducer directivities and additional information to solve the angular ambiguity.

This paper will describe firstly the microphone set-up and its associated algorithm. Secondly the performances of sound localization will be assessed.

SOURCE LOCALIZATION USING MICROPHONE DIRECTIVITY PATTERN

Microphone Array Layout

The microphone array is composed of 3 cardioid microphones (see Figure 1): the first one pointing to the x axis (right), the second one to the opposite direction (left) and the third one to the z axis (top).



a *b*

FIGURE 1 Layout of the microphone device. a) 3D array. b) 2D array.

Spatial Information Achieved from Microphone Directivity

The method operates in the time/frequency domain. It is assumed that only one sound source is present at each moment for a single frequency bin. Equations will be presented for a fixed frequency. Before applying FFT, time samples are weighted by a soft edge window to avoid oscillations in frequency domain. In terms of signal processing, the choice of the window type and its length is important. The slope and frequency bounding affect results of near frequencies. To avoid instability of the source localization, results can be smoothed frequency and time wise.

The directivity of the nth cardioid microphone is represented by

$$\mathrm{M_n}(\alpha_\mathrm{n}) = \frac{1}{2}(1 + \alpha_\mathrm{n}) \tag{1}$$

where

$$\alpha_\mathrm{n} = \overrightarrow{d_s} \cdot \overrightarrow{d_{pn}} \tag{2}$$

The vector $\overrightarrow{d_s}$ defines the source direction and the vector $\overrightarrow{d_{pn}}$ refers to the pointing direction of nth microphone.

In this case, the pointing direction $\overrightarrow{d_{pn}}$ can be expressed in the Cartesian basis $\mathcal{B}_c$ for each microphone by

$$\overrightarrow{d_{p1}} = \begin{vmatrix} 1 \\ 0, \\ 0 \end{vmatrix}_{\mathcal{B}_c} \quad \overrightarrow{d_{p2}} = \begin{vmatrix} -1 \\ 0 \\ 0 \end{vmatrix}_{\mathcal{B}_c} \quad \overrightarrow{d_{p3}} = \begin{vmatrix} 0 \\ 0, \\ 1 \end{vmatrix}_{\mathcal{B}_c} \tag{3}$$

Source location direction $\overrightarrow{d_s}$ can be expressed in the Spherical or Cartesian coordinate basis, respectively $\mathcal{B}_s$ or $\mathcal{B}_c$, by

$$\overrightarrow{d_s} = \underset{\mathcal{B}_s}{\begin{vmatrix}\theta\\ \phi\\ r\end{vmatrix}} = \underset{\mathcal{B}_c}{\begin{vmatrix}r\cos\phi\cos\theta\\ r\cos\phi\sin\theta\\ r\ \sin\phi\end{vmatrix}} \tag{4}$$

where the Spherical coordinates are defined by radius r, azimuth angle θ, and elevation angle ϕ.

The directivity functions of the three microphones are [3]:

$$M_{card1}(\theta,\phi) = \frac{1}{2}(1 + r\cos\phi\cos\theta)\,, \quad a$$
$$M_{card2}(\theta,\phi) = \frac{1}{2}(1 - r\cos\phi\cos\theta)\,, \quad b \tag{5}$$
$$M_{card3}(\theta,\phi) = \frac{1}{2}(1 + r\sin\phi) \quad c$$

Since the direction is unchanged for any value of r, radius is fixed to $r = 1$ in following expressions.

The sound source produces a signal $s_0(t)$ at the basis origin. Assuming that each microphone is located at this point, their output signals $s_{cardn}(\mathrm{t})$ are:

$$s_{cardn}(t,\theta,\phi) = M_{cardn}(\theta,\phi)s_0(t) \tag{6}$$

The signals $\mathrm{S_{cardn}(t,\theta,\phi)}$ allow to getting three data:

1) The monophonic signal of the sound source, by using relations (5).a (5).b and (6)

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi) \tag{7}$$

2) The elevation angle of the sound source, by using equations (5).c and (6)

$$\phi = \sin^{-1}\left[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1\right] \tag{8}$$

3) The azimuth angle of the sound source, by using relations (5).a (5).b

$$\theta\ = cos^{-1}\left[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2\cdot\cos\phi}\right] \tag{9}$$

Alternately it is possible to use other directivity pattern microphones to reconstruct cardioid directivity virtually (see Section: "Synthesizing Virtual Cardioid Microphones").

The same method can be used on a 2D version using only the two microphones corresponding to the horizontal plane. In this case an arbitrary elevation must be fixed. This introduces an error increasing with the angular mismatch from the chosen elevation and the real position.

In this first step, source location is calculated using exclusively the microphone directivity pattern. However the azimuth is estimated with a sign ambiguity (front-rear) due to the cosine of Equation (9). This ambiguity can be solved by moving microphones perpendicularly to their pointing direction, which will be illustrated in the next section.

Front – Rear Ambiguity Resolution Using Time Delay

If we consider now that the n[th] microphone is located at the position defined by the $\overrightarrow{E_n}$ vector:

$$\overrightarrow{E_n} = \underset{\mathcal{B}_c}{\begin{vmatrix}x_n\\ y_n\\ z_n\end{vmatrix}} \tag{10}$$

its output signal, s_{cardn}, becomes:

$$s_{cardn}(t,\theta,\phi) = M_{cardn}(\theta,\phi\ )\times s_0(t-\tau_n) \tag{11}$$

where τ_n is the delay induced by the distance between the n[th] microphone and the basis origin given by

$$\tau_n = \frac{1}{c}\,\overrightarrow{d_s}\cdot\overrightarrow{E_n} \tag{12}$$

From equation (4) the delay becomes

$$\tau_n = \frac{1}{c}\,(x_n\,r\cos\phi\cos\theta + y_n r\cos\phi\sin\theta + z_n r\ \sin\phi). \tag{13}$$

In frequency domain $s_{cardn}(t)$ becomes $S_{cardn}(\omega)$ (where $\omega = 2\pi f$ is the angular frequency with f the time frequency) by applying a Fourier Transform $FT[\ \ ]$

$$S_{cardn}(\omega,\theta,\phi)\ = FT[s_{cardn}(t,\theta,\phi\ )] = M_{cardn}(\theta,\phi)S_0(\omega)\,e^{-j\omega\tau_n}\ , \tag{14}$$

with

$$S_0(\omega) = |S_0(\omega)|\,e^{j\angle S_0(\omega)}\ . \tag{15}$$

Equation (14) becomes

$$S_{cardn}(\omega,\theta,\phi) = M_{cardn}(\theta,\phi)|S_0(\omega)|\, e^{j(\angle S_0(\omega)-\omega\tau_n)} . \quad (16)$$

Consequently

$$\angle S_{cardn}(\omega,\theta,\phi) = \angle S_0(\omega) - \omega\tau_n \quad (17)$$

and

$$\angle S_1 - \angle S_2 = -\omega(\tau_1 - \tau_2). \quad (18)$$

The time delay $\tau_{12} = \tau_1 - \tau_2$ between the two microphones is expressed by:

$$\tau_{12} = -\frac{1}{\omega}(\angle S_1 - \angle S_2). \quad (19)$$

Since this result is used to solve the ambiguity in relation (9) only its sign is needed. It is inserted in this latter as:

$$\theta = \frac{\tau_{12}}{|\tau_{12}|} cos^{-1}\left[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}\right] \quad (20)$$

SOUND LOCALIZATION USING COINCIDENT BI-DIRECTIONAL MICROPHONES

a *b*

FIGURE 2 Layout of the coincident microphone device. a) 3D array. b) 2D array.

Now it will be shown how to use the equations (7), (8) and (9) with virtual cardioid microphones synthesized from bidirectional microphones. The method is described here for an array composed of two bidirectional microphones and a cardioid one.

Microphone Layout

The array is composed of two bidirectional microphones pointing x and y axis over the horizontal plane and a third cardioid microphone pointing to the Z axis (see **FIGURE 2**). It should be noticed that alternately a soundfield® microphone [2] could be used since the X and Y components of B-format are equivalent to the bidirectional signal described above.

Synthesizing Virtual Cardioid Microphones

The signals delivered by the three microphones are [3]:

$$s_{bi1}(t,\theta,\phi) = \mathrm{s_0(t)}\,\sin\theta\, cos\phi \quad a$$

$$s_{bi2}(t,\theta,\phi) = \mathrm{s_0(t)}\,\sin\theta\, cos\phi \quad b \quad (21)$$

$$s_{card3}(t,\theta,\phi) = \frac{\mathrm{s_0(t)}}{2}(1+\sin\phi) \quad c$$

The virtual signals $s_{cardvirtn}(t,\theta,\phi)$ are obtained using the expressions:

$$s_{cardvirt1}(t,\theta,\phi) = \frac{s_0(t)}{2}\left(1+\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}\right) \quad a$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{s_0(t)}{2}\left(1-\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}\right) \quad b \quad (22)$$

where the pressure signal $s_0(t)$ is estimated by:

$$s_0(t) = \frac{s_{bi1}^2(t,\theta,\phi) + s_{2bi}^2(t,\theta,\phi) + 4s_{card3}^2(t,\theta,\phi)}{4s_{card3}(t,\theta,\phi)} \tag{23}$$

Alternately, if a Soundfield B-format signal is used

$$s_0(t) = W(t) \tag{24}$$

and

$$s_{cardvirt3}(t,\theta,\phi) = \frac{s_0(t)}{2}\left(1 + \frac{\mathrm{Z}(t,\theta,\phi)}{s_0(t)}\right) \tag{25}$$

where the signal Z(t) refers to the Z component of the B-format.

Sound Localization Using Acoustic Intensity

The signals of the virtual cardioid microphones (eq.(22)) allow to estimating the elevation and azimuth angle of the sound source using relations (8) and (9) but with a front-rear ambiguity which can not be solved by introducing a time delay since a coincident microphone array is used. Instead spatial information will be obtained from acoustic intensity.

Acoustic intensity vector is linked to the acoustic pressure and acoustic velocity by the relation [5]:

$$\vec{\boldsymbol{I}}(\omega) = \begin{vmatrix} \boldsymbol{I_x}(\omega) \\ \boldsymbol{I_y}(\omega) \\ \boldsymbol{I_z}(\omega) \end{vmatrix} = \frac{1}{2} Re\left\{ P^*(\omega) \begin{vmatrix} V_x(\omega) \\ V_y(\omega) \\ V_z(\omega) \end{vmatrix} \right\} \tag{26}$$

where $P^*(\omega)$ is the complex conjugated of the acoustic pressure and $V_x(\omega), V_y(\omega)$ et $V_z(\omega)$ are the x, y, z components of the acoustic velocity [5].

In the case of a progressive plane wave the acoustic pressure is expressed by :

$$P\left(\omega, r, \theta_r, \phi_r\right) = P_0(\omega) e^{-j\vec{k}\vec{r}} \tag{27}$$

where $\vec{k}$ is the wave vector.

Euler's equation gives the acoustic velocity as a function the acoustic pressure:

$$\vec{V}(\omega, r, \theta, \phi) = \frac{P(\omega, r, \theta, \phi)}{\rho c} \frac{\vec{k}}{|\vec{k}|} \tag{28}$$

where ρ is medium density and c the speed of sound.

Therefore acoustic intensity components are

$$\boldsymbol{I_\mu}(\omega) = \frac{|P_0(\omega)|^2}{2\rho c} \frac{k_\mu}{|\vec{k}|} \tag{29}$$

where μ represents x, y or z.

Thus it is observed that the acoustic intensity vector has the same direction as $\vec{k}$, and can therefore be used to estimate the direction of the sound source.

Bidirectional coincident array deliver pressure gradient information which leads directly to the acoustic velocity. For instance, the horizontal plane projection of velocity is given by:

$$V_{xy}(\omega) = \frac{1}{\rho c}\left[S_{bi1}(\omega)\overrightarrow{e_x} + S_{bi2}(\omega)\overrightarrow{e_y}\right] \tag{30}$$

From Equation (26), the acoustic intensity components are:

$$\boldsymbol{I_x}(\omega) = \frac{1}{2\rho c} Re[S_0^*(\omega) S_{bi1}(\omega)] \quad a$$
$$\boldsymbol{I_y}(\omega) = \frac{1}{2\rho c} Re[S_0^*(\omega) S_{bi2}(\omega)] \quad b \tag{31}$$

Elevation and azimuth angle are then obtained from Equation (29) [6]:

$$\theta(\omega) = \tan^{-1}\left(\frac{\boldsymbol{I_y}(\omega)}{\boldsymbol{I_x}(\omega)}\right) \quad a \tag{32}$$

$$\phi(\omega) = \tan^{-1}\left(\frac{\boldsymbol{I_z}(\omega)}{\sqrt{\boldsymbol{I_x^2}(\omega) + \boldsymbol{I_y^2}(\omega)}}\right) \qquad b$$

Solving Ambiguity Localization for Coincident Arrays

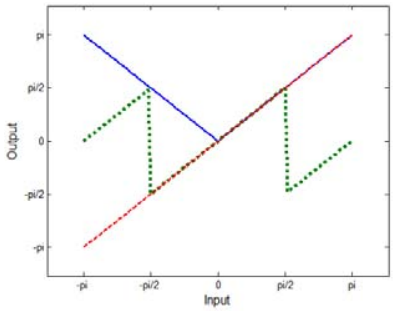


FIGURE 3 Angle estimation ambiguity: Cardioid directivity method (continuous blue line; see Equation (9)), Acoustic intensity method (dotted green line; see Equation (32)), Theoretical position (red dashed line)

TABLE 1 Ambiguity resolution by cross-checking the angular estimation from the directivity (Equation (9)) and intensity (Equation (32)) methods

real $\boldsymbol{\theta}$	estimated $\boldsymbol{\theta}$		operation to solve ambiguity	
	Directivity	Intensity	Directivity	Intensity
Case 1 $:\boldsymbol{\theta} \in \left[-\pi, -\frac{\pi}{2}\right]$	$\boldsymbol{\theta_D} \in \left[\frac{\pi}{2}, \pi\right]$	$\boldsymbol{\theta_I} \in \left[0, \frac{\pi}{2},\right]$	$-\boldsymbol{\theta_D}$	$\boldsymbol{\theta_I} - \boldsymbol{\pi}$
Case 2 $:\boldsymbol{\theta} \in \left[-\frac{\pi}{2}, 0\right]$	$\boldsymbol{\theta_D} \in \left[0, \frac{\pi}{2},\right]$	$\boldsymbol{\theta_I} \in \left[-\frac{\pi}{2}, 0\right]$	$-\boldsymbol{\theta_D}$	$\boldsymbol{\theta_I}$
Case 3 $:\boldsymbol{\theta} \in \left[0, \frac{\pi}{2},\right]$	$\boldsymbol{\theta_D} \in \left[0, \frac{\pi}{2},\right]$	$\boldsymbol{\theta_I} \in \left[0, \frac{\pi}{2},\right]$	$\boldsymbol{\theta_D}$	$\boldsymbol{\theta_I}$
Case 4 $:\boldsymbol{\theta} \in \left[\frac{\pi}{2}, \pi\right]$	$\boldsymbol{\theta_D} \in \left[\frac{\pi}{2}, \pi\right]$	$\boldsymbol{\theta_I} \in \left[-\frac{\pi}{2}, 0\right]$	$\boldsymbol{\theta_D}$	$\boldsymbol{\theta_I} + \boldsymbol{\pi}$

Localization based on the cardioid directivity allows to obtaining the azimuth angle with front – rear ambiguity due to the cosine relation involved in its estimation (eq.(9)). This estimation can be solved for non coincident arrays by inserting a delay. For coincident arrays it is possible to use acoustic intensity to estimate azimuth angle, but this time with left - right ambiguity due to the inverse of the tangent in the relation (32).a.

As shown by **FIGURE 3** and **TABLE 1**, the front - rear ambiguity is complementary to the left - right ambiguity. Four cases are pointed out, corresponding to the four combinations of the two ambiguous estimations. The real position can then be found using a conditional research.

In theory once the ambiguity is solved, both methods (i.e Equation (9) or Equation (32)) give the same result. However they may be slightly different in practice. Depending on the sound scene, the sound stimulus or the noise level, one method can achieve better performances.

LOCALIZATION ASSESSMENT

A computer program stimulates the signals which would have been recorded by the two microphone array set-ups previously described. Various *stimuli* were used (music, random noise, noise band and harmonic tone).

Evaluation Criteria

The azimuth and elevation error E_θ and E_ϕ are calculated here as the angular distance between the real location and the estimated one. The total error E_t is the angular distance between the real and the estimated location on the sphere (see eq(33)). The angular distance is calculated using the scalar product of the real and estimated position.

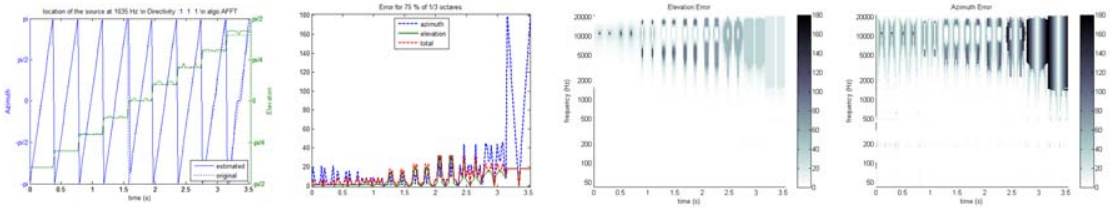
$$E = \cos^{-1}\left(\frac{\overrightarrow{V_{real}} \cdot \overrightarrow{V_{est}}}{\left\|\overrightarrow{V_{real}}\right\|\left\|\overrightarrow{V_{est}}\right\|}\right) \tag{33}$$

where

$$\overrightarrow{V_{real}} = \begin{matrix}\\ \mathcal{B}_s\end{matrix}\begin{vmatrix}\theta_{real}\\ \phi_{real}\\ 1\end{vmatrix}, \overrightarrow{V_{est}} = \begin{matrix}\\ \mathcal{B}_s\end{matrix}\begin{vmatrix}\theta_{est}\\ \phi_{est},\\ 1\end{vmatrix} \tag{34}$$

When E_θ or E_ϕ are calculated, ϕ_{real} and ϕ_{est}, or θ_{real} and θ_{est} component are fixed to 0 respectively. The error is expressed in degrees where 0° states the best estimation and 180° the worst one. In addition a new criterion is proposed, computed as the error level obtained by at least 75% of 1/3 octave spectrum. It will be referred to as E_{75}.

Results

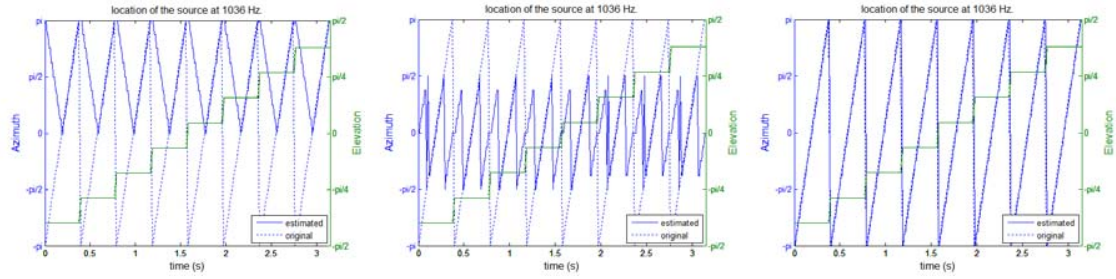


a *b* *c* *d*

FIGURE 4 Source localization of a random noise moving around from bottom to the top of a 3 cardioid microphones array. Horizontal plane microphones are separated by 2cm.

a) Source location at 1036 Hz, azimuth (blue) and elevation (green). b) E_{75} criterion, azimuth (blue), elevation (green), total (red) c) Elevation source location error. d) Azimuth localization error.

FIGURE 4 depicts the results obtained when localizing a random noise moving around and from bottom to the top. Localization accuracy is affected by the microphone spacing because the reconstruction of the omnidirectional pressure S_0 is altered. As shown in FIGURE 4c, high frequencies are more affected when the wavelength is closer to the microphone distance. For high elevations variations of the cardioid pattern (s_{card3}) are slow which results in a poor angular discrimination. As azimuth localization uses elevation (cf. eq. (20)), azimuth estimation is consequently degraded, as shown by FIGURE 4d.

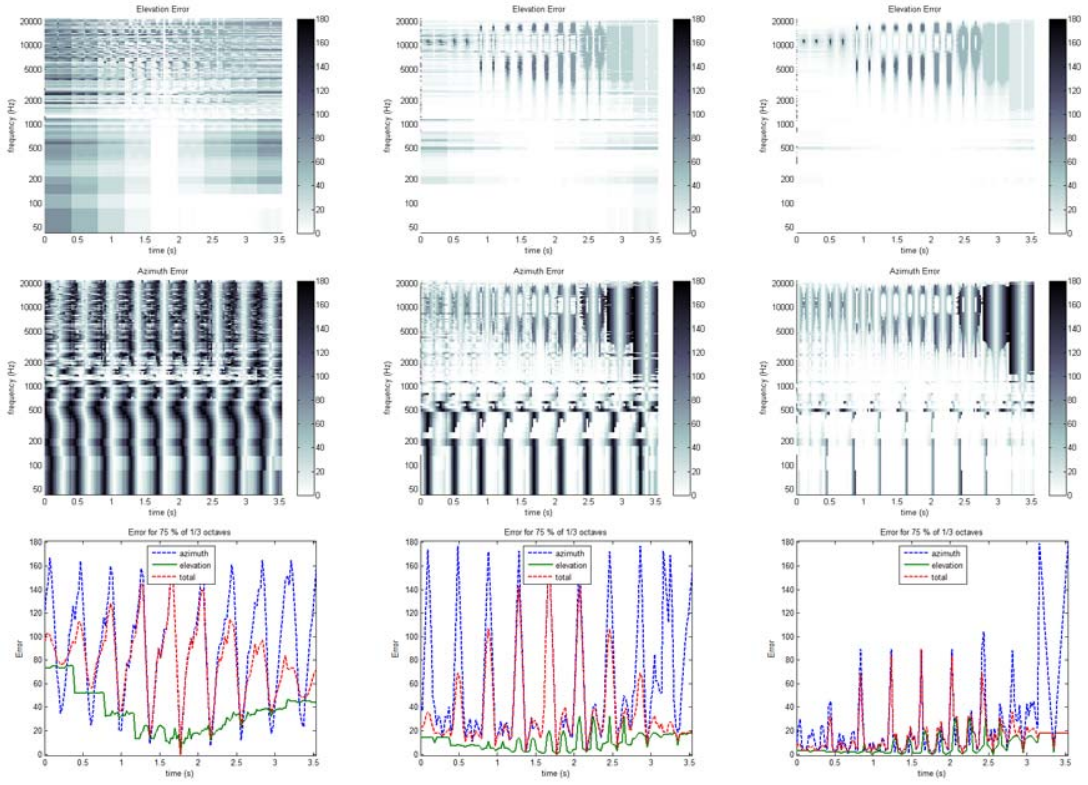


A *B* *c*

FIGURE 5 Source localization of a random noise in azimuth (blue) and elevation (green) picked up by 2 bidirectional + 1cardioid microphone array at 1036 Hz found with a). directivity only, b) intensity only, c) directivity and intensity

When coincident arrays are used, estimation of the pressure signal is more accurate (see eq.(23)) and results are not affected with elevation over all frequencies (see FIGURE 5). However, in practice coincident microphone arrays are impossible to build. As a consequence some artifacts will always occur in high frequencies.

As it has been specified before, it is assumed that only one source is present at each moment and at each frequency. In practice, acoustic field is complex and more than one source is present affecting localization accuracy (see **FIGURE 6** and **FIGURE 7**). Low energy signals close to high energy signals are then localized to the direction of the higher one. When the target sound source level is higher than 12 dB to disturbing noise, localization error is



A *B* *c*

FIGURE 6 Source localization evaluation of a random noise: Elevation error (1st row), Azimuth error (2nd row) and E_{75} criterion (3rd row) picked up with a 3 cardioid microphone array in presence of a disturbing second noise source at θ=0 and ϕ=0 and a level difference of a). 0 dB, b) -12 dB, c) -20 dB

weak: Some front-rear confusions are introduced below 500 Hz since phase information of the target source is altered by disturbing second source. Obviously this is only observed in the case of the cardioid array. The influence of a disturbing source becomes insignificant for level differences higher than 20 dB.

Contrasting with **FIGURE 6**, it is observed for the bidirectional array that error increases unexpectedly for low elevations (see **FIGURE 7**-3c). Indeed the target source level is disadvantaged by the array directivity which is "deaf" at those directions. This phenomenon can be turned into advantage if the disturbing noise is placed at the "deaf" area of the array. On the contrary for cardioid array the target source is homogenously picked-up at all directions. As suggested by **FIGURE 6**, azimuth error E_θ has not the same impact in terms of total angular distances E_{total}in function of the elevation angle. **FIGURE 7** clearly shows that azimuth error has less impact when the source is near to the poles.

CONCLUSION AND FUTURE WORKS

In order to provide 3D audio tools for consumer device, we propose a recording solution using a small number of microphones. The omnidirectional pressure signal is recomposed and compared with the directional microphone output. Sound source location is estimated in azimuth and elevation using microphone directivity. However the localization is front-rear ambiguous. For non-coincident arrays, this ambiguity is solved by time difference between microphones, whereas for coincident arrays, the acoustic intensity vector is used. In the case of non-coincident arrays, the omnidirectional pressure component is not properly reconstructed at wavelengths closer to microphone spacing, which alters the localization.

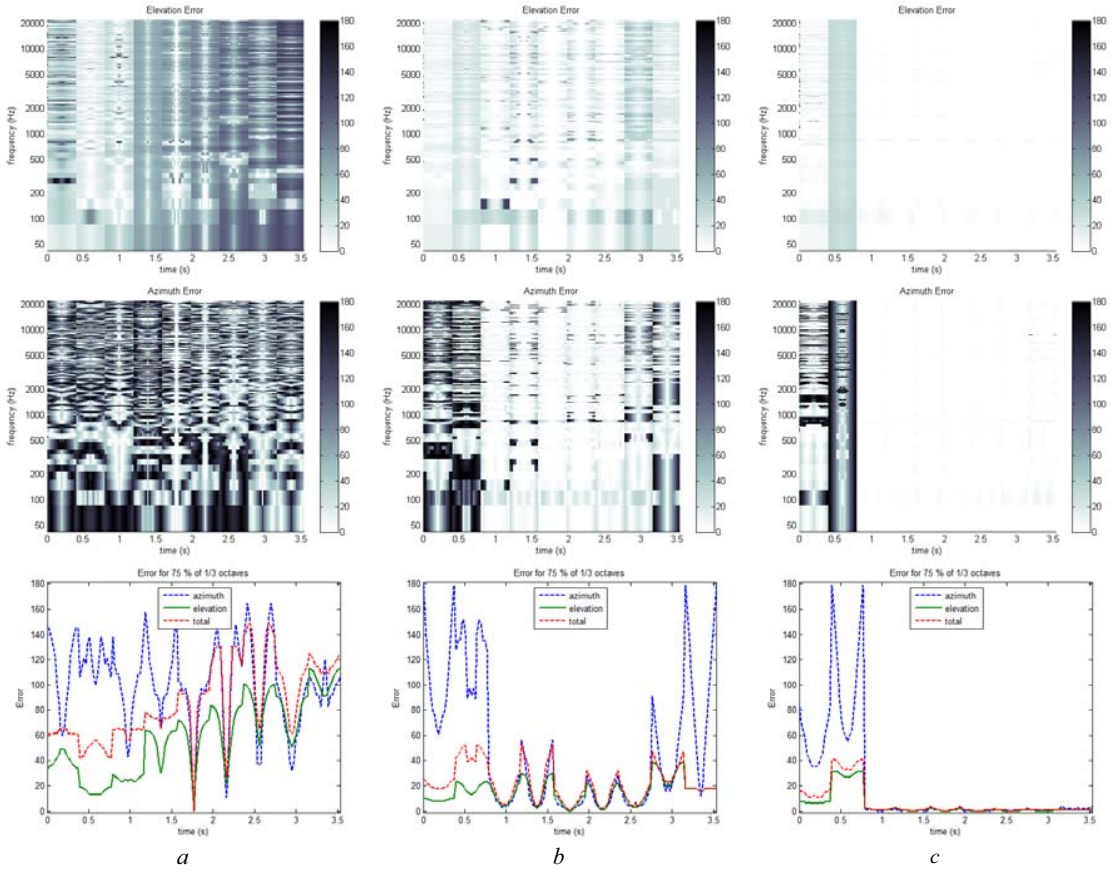


FIGURE 7 Source localization evaluation of a random noise : Elevation error (1st row), azimuth error (2nd row) and E_{75} criterion (3rd row) picked up with a 2 bidirectional + 1 cardioid microphone array in presence of a disturbing second noise source at θ=0 and ϕ=0 and a level difference of a). 0 dB, b) -12 dB, c) -20 dB

Sound scene analysis presented in this paper can be used as the first step of object based spatial audio representation. Using the information of source position, it is possible to render the sound scene over any type of spatial audio system such as stereo, 5.1, 7.1, 22.2, Higher Order Ambisonics [7] or Wave Field Synthesis [8].

REFERENCES

[1]. J. Sunier, The story of stereo: 1881-. Gernsback Library, 1960.
[2]. R. Nicol, « Représentation et perception des espaces auditifs virtuels », HDR, Université du Maine, Le Mans, France, 2010.
[3]. J. Jouhaneau, Notions élémentaires d'acoustique: Électroacoustique. Tec & Doc Lavoisier, 1999.
[4]. Michel A. Gerzon et Peter G. Craven, « Coincident microphone simulation covering three dimensional space and yelding various directional outputs », U.S. Patent 4,042,77916-août-1977.
[5]. M. Bruneau, Manuel d'acoustique fondamentale. Hermès, 1998.
[6]. V. Pulkki, « Directional audio coding in spatial sound reproduction and stereo upmixing », in Proc. of the AES 28th Int. Conf, Pitea, Sweden, 2006.
[7]. J. Daniel, « Evolving views on HOA: From technological to pragmatic concerns », Ambisonics Symposium 2009, June 25-27, Graz, 2009.
[8]. A.J. Berkhout, D. de Vries & P. Vogel, « Acoustic Control by Wave Field Synthesis », J. Acoust. Soc. Am., 1993, 93, pp. 2764-2778.

E.4 A Surround Microphone in Your Pocket [Palacino and Nicol, 2013c]

ASA Lay Language Papers
165th Acoustical Society of America Meeting

[Lay Language Paper Index | Press Room]

A Surround Microphone in Your Pocket

Julian Palacino – julianpalacino@hotmail.com
Rozenn Nicol – rozenn.nicol@orange.com
Orange Labs
SVQ/TPS
2 Av Pierre Marzin
22307 Lannion

Popular version of paper 4aSP2
Presented Thursday morning, June 6, 2013
ICA 2013 Montreal

Nowadays, everyone has heard about 3D video but only few of us have already heard 3D audio. The aim of 3D audio techniques is to record and reproduce sound with the naturalness as we perceive it in real life.

Human beings use binaural perception (using both ears) to recognize the position of an acoustic source. A sound arrives first and louder to the ear closest to the sound source and our head and pine modifies sounds in function of its direction. Our brain learns to interpret those cues and let us determine the position of a sound.

For several decades spatial or 3D audio has been only used by movie makers, music composers and researchers in laboratories; but because of its complexity, the general public hasn't been concerned about those techniques. In addition, dedicated devices such as microphones and loudspeakers are expensive and cannot be used without some expertise of audio capturing and reproduction.

Currently, audio techniques such stereo, 5.1 or 7.1 gives a limited spatial impression. Naturalness is sacrificed to get a good quality and resulting sound is different as it was listened during the performance.
(see figure 1- Traditional recording process and downmix: Instruments, sound scene and ambience are generally picked up with a big number of microphones. All those signals are then mixed by a sound engineer who sets up how loud each microphone is played on each loudspeaker).

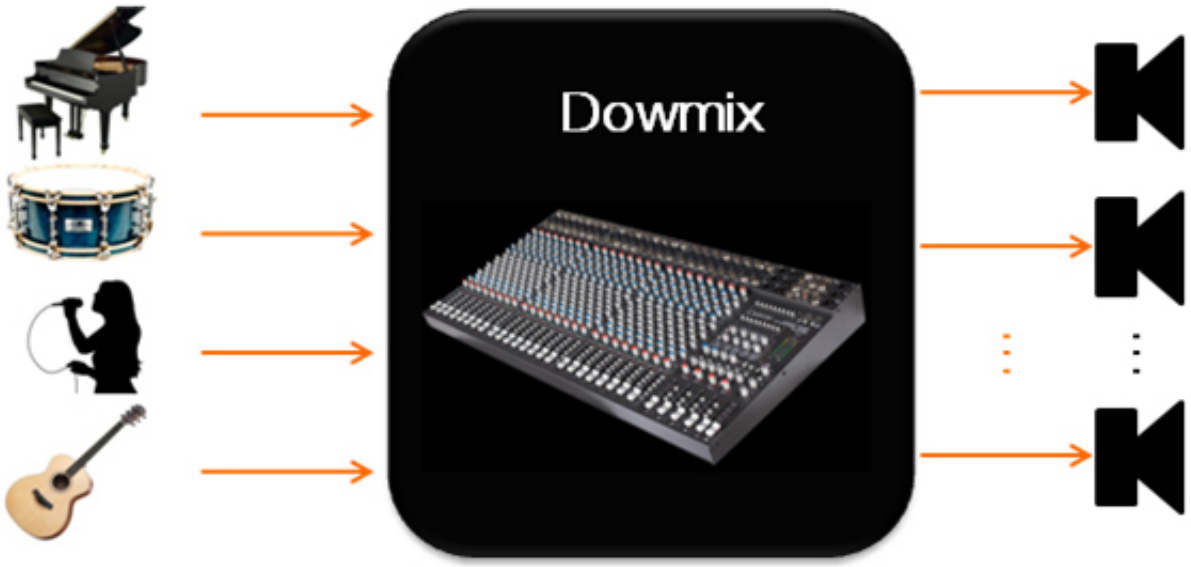


Figure 1

Other approaches have also been used in order to pick up and reproduce the acoustic waves as close as possible to the sound produced during the performance. Binaural uses a dummy head equipped of two microphones in the ears, audio signals picked up are modeled by the shape of the head and the ears pine [2] to reproduce a sound as close as possible to the one heard by a human (see figure 2, Neuman KU4 – Comercial dummy head for binaural recordings). For the restitution, the sounds must be reproduced close to the listener ears which suppose the use of headphones. Ambisonics and higher order ambisonics (HOA) allow to record a 3D acoustic image using a big number of microphones (See figure 3, Eigenmike – HOA microphone array composed by 32 microphones). For the reproduction, this image is projected over a big number of loudspeakers [5][8][9].

Figure 2

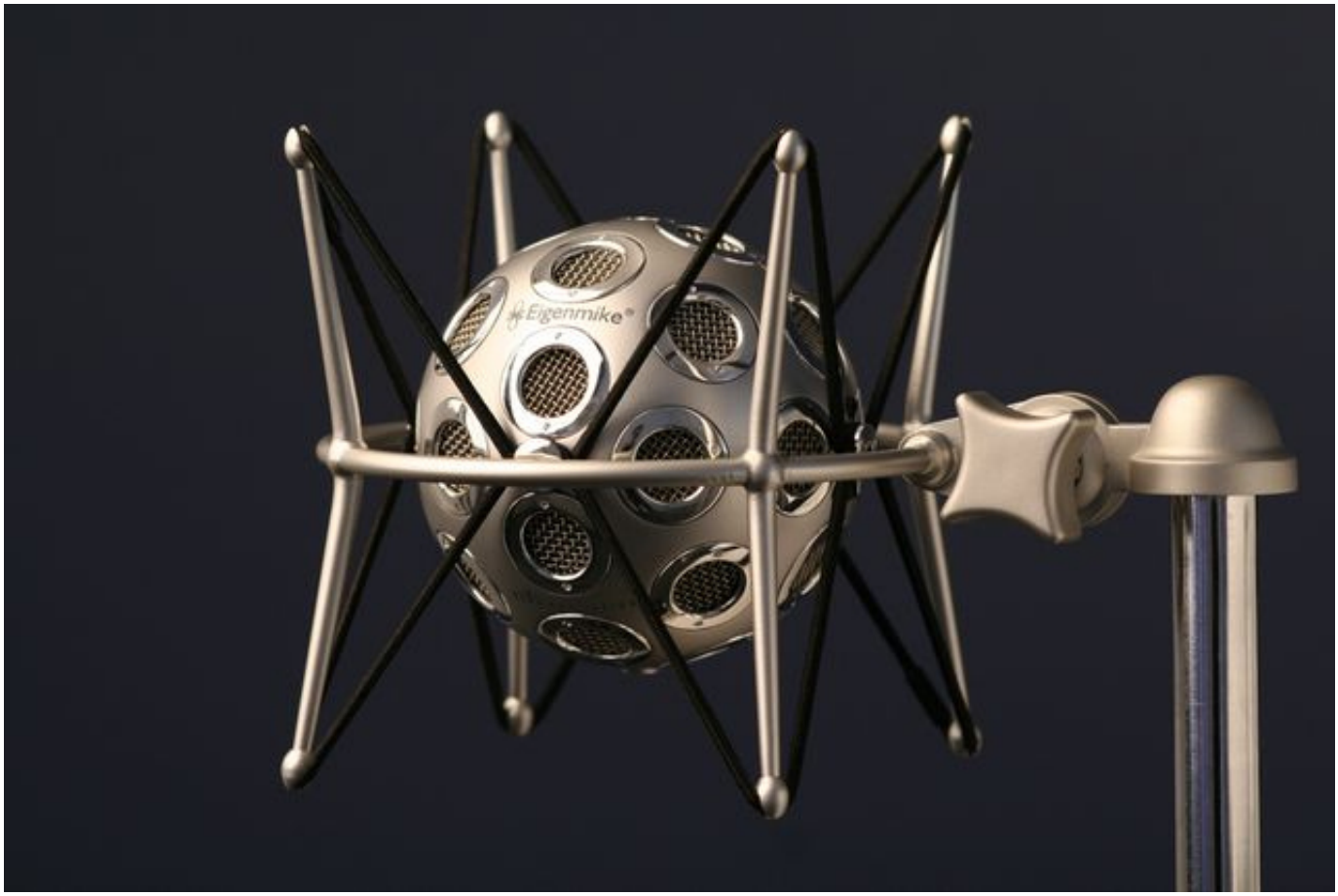


Figure 3

Nowadays, the main barrier preventing a consumer solution from capturing spatial audio is the big number of microphones and loudspeakers needed to get an accurate 3D sound image or the size of the recording devices. In order to break down this barrier, we propose a new 3D audio recording set-up which is composed of a three microphone array capable of getting the full 3D audio information.

Microphone array is composed of three cardioid microphones (See figure 4 Layout of the microphone device), one pointing left, the other pointing right and the third one pointing upwards. In order to introduce a time delay between them, the horizontal microphones are misaligned.

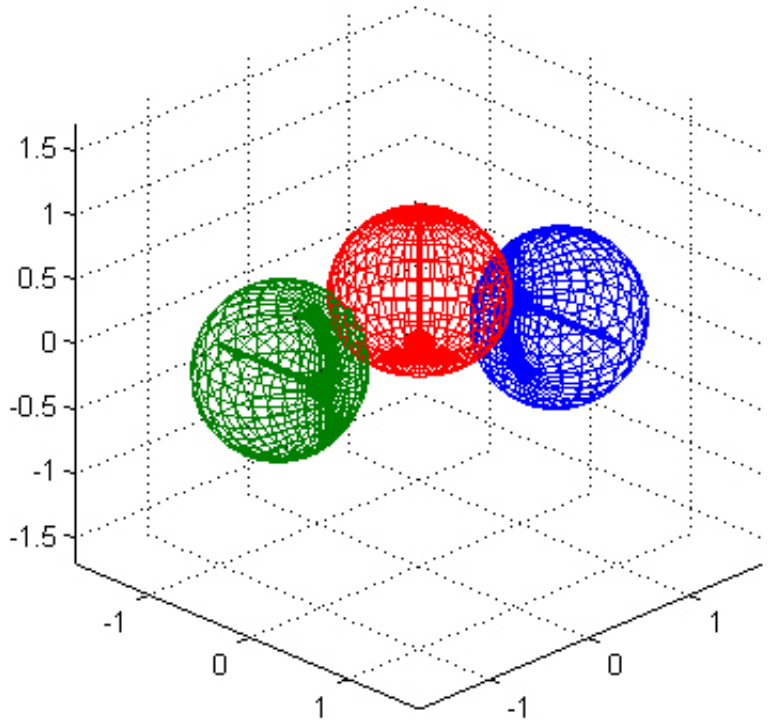


Figure 4

Contrary to a sound picked up with an omnidirectional microphone which the level doesn't depend on the direction of the source, a sound picked up with a cardioid one has a particular level corresponding to the direction of the sound source [4].

In our case, it is possible to recover a sound signal close to the one picked up by an omnidirectional microphone adding signals coming from the cardioids microphones pointing left and right. Comparing the level of this reference signal with the signal level picked up by each microphone, one can get the position of the sound source. As the array is symmetrical, results are front-rear ambiguous. As the sound arrives first to the microphone closest to the source this information is also used to solve the ambiguity and get the right position.

The two microphones placed over the horizontal plane give the localization information in terms of azimuth and the one pointing upwards the elevation position. Those two parameters give the right location of the sound source. In several cases, elevation information is not needed and only a two microphones array gives the necessary information.

As sound sources are well localized this technique can be the first step for a sound reproduction over any kind of spatial audio system such as stereo, 5.1, 22.1, binaural, ambisonics or wavefield synthesis. The number of microphones and its size make this system compatible with mobile devices such as smartphones and tablets and we can expect to find this system in the market as a third-party accessory for audio applications or directly embedded in this kind of devices.

References

1. J. Sunier, The story of stereo: 1881-. Gernsback Library, 1960.
2. J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization. Mit Press, 1997.
3. R. Nicol, « Représentation et perception des espaces auditifs virtuels », HDR, Université du Maine, Le Mans, France, 2010.
4. J. Jouhaneau, Notions élémentaires d'acoustique: Électroacoustique. Tec & Doc Lavoisier, 1999.

5. Michel A. Gerzon et Peter G. Craven, « Coincident microphone simulation covering three dimensional space and yelding various directional outputs », U.S. Patent 4,042,77916-août-1977.
6. J. Daniel, « Evolving views on HOA: From technological to pragmatic concerns », Ambisonics Symposium 2009, June 25-27, Graz, 2009.
7. A.J. Berkhout, D. de Vries & P. Vogel, « Acoustic Control by Wave Field Synthesis », J. Acoust. Soc. Am., 1993, 93, pp. 2764-2778.

[Lay Language Papers Index | Press Room]

E.5 Des HRTF aux Object-RTF : Système de prise de son 3D pour dispositifs nomades [Palacino and Nicol, 2014]

Des HRTF aux Object-RTF : Système de prise de son 3D pour dispositifs nomades

J. Palacino et R. Nicol
Orange Labs, 2 Avenue Pierre Marzin, 22300 Lannion, France
julianpalacino@gmail.com

Les systèmes de prise de son 3D disponibles aujourd'hui reposent sur 2 principales stratégies : soit la captation de tout ou partie des indices naturels de localisation auditive (stéréophonie ou technologie binaurale), soit la décomposition spatiale de l'onde acoustique sur une base de fonctions propres à partir d'un réseau microphonique (Higher Order Ambisonics par exemple). Une approche plus récente propose d'extraire directement l'information de direction des sources sonores à partir d'une prise de son par un microphone ambisonique à l'ordre 1 (Pulkki [14] et Berge [1]) permettant en même temps, la compression de l'information et la restitution spatiale. L'extraction de la direction est basée sur les propriétés du format B et les descripteurs de localisation qui lui sont associés [3, 4]. Dans un autre but, Keyrouz [5, 6] et MacDonald [8] ont présenté une méthode qui exploite l'information spatiale contenue dans les signaux binauraux en la comparant avec l'ensemble d'une base de HRTF (Head Related Transfer Function) pour identifier la direction d'une source sonore. Nous proposons ici une nouvelle approche de prise de son 3D inspirée du procédé de Pulkki et Berge, mais utilisant une procédure de localisation proche de celle de Keyrouz et MacDonald. Plus précisément, nous généralisons cette dernière en utilisant un dispositif de captation constitué de deux ou trois microphones intégrés à un objet diffractant dont on définit les fonctions de transfert directionnelles ou "Object-RTF" (Objet Related Transfert Function) à l'image des HRTF. Après une description du procédé, ses performances de localisation sont évaluées sur une grille de critères objectifs qui ont été introduits dans [11].

1 Introduction

Il existe aujourd'hui plusieurs méthodes pour capter l'information spatiale d'une scène sonore. Néanmoins, à cause de leur complexité de mise en œuvre et de la taille des dispositifs, ces technologies n'ont pas atteint le grand public. Afin de favoriser leur dissémination, l'étude présentée ici s'inscrit dans des travaux visant à doter les terminaux mobiles de solutions adaptées de prise de son 3D. Cet objectif impose des contraintes d'encombrement et de mobilité qui sont liées à ce type de terminaux.

Les systèmes de prise de son 3D disponibles aujourd'hui reposent sur trois principales stratégies. La première s'intéresse à la captation et la reproduction des indices de localisation naturelle. Un exemple est la technologie binaurale qui consiste à capter les signaux acoustiques à l'entrée du conduit auditif. La prise de son est couramment réalisée par une tête artificielle. La seconde méthode vise la décomposition physique des propriétés spatiales du champ acoustique, par exemple sur une base de fonctions propres. Il s'agit notamment de la technologie ambisonique qui repose sur des réseaux sphériques de microphones. Une troisième approche propose d'extraire pour chaque source simplement deux informations : la direction et un signal représentatif de la source sonore, à partir d'une prise de son par un dispositif microphonique donné [1] [14].

Cette dernière solution, qui s'apparente à une représentation « objet » de la scène sonore, offre deux avantages : d'une part il s'agit d'une représentation efficace en termes de compression de l'information, d'autre part elle est compatible avec différents formats de reproduction dans la mesure où, grâce à l'information de localisation des sources, il suffit de synthétiser des sources virtuelles selon le format correspondant au système de restitution disponible.

Dans notre contexte d'application aux terminaux mobiles, les deux premières stratégies de prise de son 3D sont inadaptées au vu des contraintes d'encombrement et de mobilité. Nous avons donc opté par la troisième méthode, en imposant comme contrainte supplémentaire l'utilisation d'un dispositif de captation à faible nombre de capteurs. Nous avons étudié comment un dispositif microphonique composé de trois capteurs cardioïdes est capable de capter l'intégralité de l'information spatiale de la scène sonore. En aval de la prise de son, un algorithme de post-traitement spatial permet d'extraire, en tirant parti à la fois de la directivité et des différences de temps, d'une part les paramètres de localisation de sources sonores et d'autre part, un signal de référence représentatif du signal acoustique émis par les sources.

Dans une première étape, les microphones ont été considérés comme de capteurs idéaux et l'influence acoustique du support microphonique n'a pas été prise en compte. Dans cet article, nous décrivons comment le processus de localisation de sources peut être adapté pour intégrer les caractéristiques réelles des microphones, incluant potentiellement l'influence du support microphonique et du corps du terminal. Ces caractéristiques sont décrites sous la forme de fonctions de transfert dites « Object-RTF » constituant une généralisation du concept de HRTF (Head Related Transfer Function).

Le dispositif de prise de son est d'abord brièvement rappelé. Ensuite, le concept d'Object-RTF est introduit. Puis, son utilisation pour la localisation des sources est décrite. Pour l'identification de la direction, trois critères pour la mesure de distance sont proposés. Leur intérêt est comparé dans la dernière partie et les performances de l'algorithme de localisation sont évaluées en termes d'erreur angulaire et de complexité algorithmique.

2 Travaux antérieurs

Afin de proposer une méthode permettant l'enregistrement d'une scène spatiale deux méthodes analogues aux techniques de Berge [1] et Pulkki [14] ont été proposés dans [10, 11]. Ces méthodes cherchent principalement à représenter une scène sonore complexe, grâce à l'extraction des positions de sources la composant et d'un signal décrivant l'ensemble de ces sources.

Les méthodes proposées sont associées à un dispositif microphonique qui par ces caractéristiques remplit le cahier de charges pour une intégration en mobilité, à

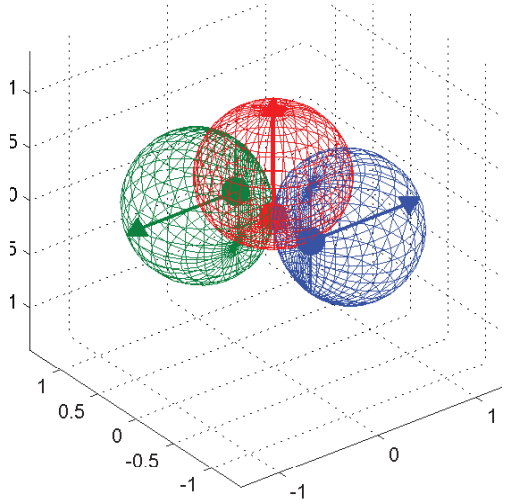


FIGURE 1 – Configuration microphonique.

savoir sa faible taille et son faible nombre de capteurs.

Le dispositif microphonique est composé de trois capteurs cardioïdes (cf. figure 1). Le premier microphone pointe vers les x positifs (droite), le deuxième vers les x négatifs (gauche) et le troisième vers l'axe z (haut). Les deux microphones sur le plan horizontal sont écartés de 1 cm du centre selon l'axe y. Ce dispositif permet l'obtention aisée du signal représentatif de la scène sonore et la localisation de la direction des sources sur toute la sphère 3D avec uniquement 3 capteurs. Dans le cas d'une prise de son pour une diffusion sur un système multicanal classique (5.1, 7.1) ou pour une application de téléconférence, où toutes les sources se trouvent a priori sur le même plan horizontal, l'élévation n'est plus nécessaire et le système peut se réduire aux seuls deux microphones placés sur le plan horizontal.

2.1 Utilisation de la directivité pour l'extraction du signal de référence

Nous considérons le signal de pression acoustique S_o au centre du dispositif microphonique et en l'absence de celui-ci. S_o est suffisant pour caractériser complètement la scène sonore. A condition d'utiliser un dispositif comme celui décrit précédemment, une estimation de S_o peut être obtenue grâce à la directivité de microphones utilisés.

La directivité d'un microphone cardioïde est exprimée par la relation

$$M = \frac{1}{2}(1 + \vec{d_s} \cdot \vec{d_p}) \tag{1}$$

où $\vec{d_s}$ est le vecteur définissant la direction de la source et $\vec{d_p}$ est le vecteur déterminant la direction de pointage du microphone. Dans le cas présent, les directions de pointage $\vec{d_p}n$ des trois microphones peuvent être exprimées dans la base de coordonnées cartésiennes $\mathfrak{B}_c$

$$\vec{d_{p_1}} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}_{\mathfrak{B}_c} \quad \vec{d_{p_2}} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}_{\mathfrak{B}_c} \quad \vec{d_{p_3}} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}_{\mathfrak{B}_c} \tag{2}$$

et considérant que la direction de la source est exprimée dans la base $\mathfrak{B}_s$ en coordonnées sphériques ou dans la base de coordonnées cartésiennes $\mathfrak{B}_c$

$$\vec{d_s} = \begin{pmatrix} \theta \\ \phi \\ r \end{pmatrix}_{\mathfrak{B}_s} = \begin{pmatrix} r\cos\theta\cos\phi \\ r\sin\theta\cos\phi \\ r\ \sin\phi \end{pmatrix}_{\mathfrak{B}_c} \tag{3}$$

les relations de directivité son définies par :

$$M_1 = \frac{1}{2}(1 + r\cos\phi\cos\theta), \tag{4a}$$

$$M_2 = \frac{1}{2}(1 - r\cos\phi\cos\theta), \tag{4b}$$

$$M_2 = \frac{1}{2}(1 + r\sin\phi). \tag{4c}$$

Pour alléger les notations, il est possible de définir $r = 1$ car ceci ne modifie pas la direction de pointage. La source dont la direction est pointée par le vecteur $\vec{d_s}$ génère un signal S_0 à l'origine du repère. Considérant que les trois microphones y sont placés, les signaux S_n issus de chacune des capsules correspondent au signal S_0 pondéré par la fonction de directivité associée

$$S_n = M_n S_0. \tag{5}$$

Compte tenu des relations (4), le signal de référence S_0 cherché s'obtient :

$$S_0 = S_1 + S_2. \tag{6}$$

Ce signal pourra être utilisé comme porteur de toute l'information sonore de la scène analysée.

2.2 Localisation des sources basée sur la directivité de microphones

Dans [12] nous avons proposé d'utiliser la directivité des capteurs afin de localiser la source. Dans l'équation (5), le terme M qui définit la directivité est une pondération directionnelle du signal S_o. M dépend de la position de la source en coordonées sphériques θ et ϕ et est propre au microphone. Connaissant donc S_o et les fonctions directionnelles $M(\theta, \phi)$, il ne reste qu'à résoudre une équation à deux inconnues pour connaître les coordonnées de la direction de la source. La position obtenue avec cette méthode présente une ambiguïté avant-arrière liée à l'inversion d'une fonction cosinus.

Dans [11], la méthode est étendue en considérant aussi le cas des signaux ambisoniques à l'ordre 1. Les signaux du format B de l'ambisonique permettent de reconstruire virtuellement un dispositif présentant les caractéristiques du capteur décrit précédemment. Dans ce cas, étant tributaire du format d'entrée, le microphone obtenu virtuellement est coïncidant et la même ambiguïté avant-arrière est rencontrée.

2.3 Localisation des sources basée sur le vecteur intensité acoustique

Le vecteur intensité acoustique est colinéaire avec la direction de propagation de l'onde acoustique. Par extension, ce vecteur pointe la position de la source qui la génère et peut être dérivé du format

B de l'ambisonique. Ces informations permettent la localisation de la source selon la méthode [14]. Les résultats obtenus par cette technique présentent une ambigüité droite-gauche liée à l'inversion d'une tangente.

2.4 Résolution de l'ambiguïté

Afin de localiser la direction de la source sur toute la sphère 3D, il est nécessaire de résoudre les ambiguïtés liées aux techniques décrites. Dans le premier cas, présenté dans le paragraphe 2.2, l'ambiguïté de localisation est résolue par la dissymétrie sur l'axe y du dispositif en exploitant la différence de phase entre les signaux captés par les deux microphones placés sur le plan horizontal.

Dans le cas où les signaux du format B de l'ambisonique à l'ordre 1 sont utilisés, il est nécessaire de conjuguer les deux méthodes décrites précédemment. Les deux résultats présentent des ambiguïtés qui s'avèrent complémentaires. Il est possible, en les comparant, de trouver une solution unique et sans ambiguïté.

3 Les Object-RTF

Les HRTF ou fonctions de transfert liées à la tête sont les réponses en fréquence directionnelles décrivant le trajet acoustique entre une source sonore et l'entrée du conduit auditif en fonction de la direction de la provenance de l'onde. Le système auditif exploite les indices spectraux engendrés par les diffractions sur la tête, le torse et les pavillons des oreilles ainsi que les différences interaurales pour déterminer la position d'une source. Tout microphone présente une directivité ainsi qu'une réponse en fréquence intrinsèque à ses dimensions et à sa conception. Le champ sonore capté par celui-ci est aussi modifié par les diffractions et réflexions engendrées par les objets qui l'entourent, tels que les préamplificateurs ou leur support. Lorsque le microphone est inséré ou intégré à un objet, sa réponse en fréquence et sa directivité sont modifiées et peuvent être mesurées. Pour généraliser le concept d'HRTF, nous définissons les Object-RTF (*Objet Related Transfert Function*) comme les réponses en fréquence d'un microphone et dépendant de la fréquence et de la direction de la source en azimut θ et en élévation ϕ. Les Object-RTF peuvent être mesurées de manière similaire aux HRTF en utilisant des méthodes existantes [7, 9, 13]. Il s'agit principalement de mesurer la réponse du dispositif pour un ensemble de directions situées sur la sphère 3D. Dans le cas d'un microphone directionnel parfait, sa réponse est identique autour de son axe de révolution. C'est-à-dire que le signal émis par une source sonore est identique quelle que soit sa position sur le cercle défini autour de l'axe de révolution du microphone. Connaissant la directivité du microphone, ainsi que le signal capté par un microphone omnidirectionnel au même point (signal de référence), il est possible de déterminer la position du cercle contenant la source. Si un deuxième microphone est utilisé, la solution se limite à deux points à l'intersection de deux cercles. Une fois que le dispositif est inséré dans un objet, sa directivité et sa réponse en fréquence sont modifiées, on parle alors d'Object-RTF. En fonction de la géométrie de l'objet, la symétrie de la directivité des microphones est conservée ou non. Dans le cas où la géométrie est cassée, connaissant les Object-RTF ainsi que le signal de référence, l'analyse permet de déterminer une solution unique. Ainsi, l'algorithme de localisation proposé par MacDonald [8] et Keyrous [5, 6] montre la possibilité d'utiliser un jeu de HRTF afin d'obtenir une bonne localisation sans aucune connaissance a priori de la source à localiser utilisant un set de HRTF.

4 Analyse de scène sonore basée sur les Object-RTF

L'algorithme de localisation de sources basée sur les Object-RTF est la généralisation de l'utilisation des HRTF pour la localisation de sources tel qu'elle a été définie par MacDonald [8] et Keyrous [5, 6].

4.1 Les HRTF dans la localisation des sources

Les algorithmes proposés par MacDonald [8] et Keyrous [5, 6] sont utilisés pour la localisation des sources sonores à l'aide des HRTF.

Dans le domaine fréquentiel le signal issu du microphone placé à l'oreille gauche est caractérisé par :

$$S_L(f,\hat{\theta},\hat{\phi}) = S_o(f)\ HRTF_L(f,\hat{\theta},\hat{\phi}) \tag{7}$$

où $S_o(t)$ est la réponse en fréquence du signal de pression acoustique pouvant être mesuré entre les deux oreilles et en absence de la tête, et $HRTF_L$ la réponse en fréquence de l'oreille gauche pour une source placée à $(\hat{\theta},\hat{\phi})$. Par la suite, les indices L et R désignent les grandeurs correspondant aux oreilles gauche et droite respectivement.

Nous constatons qu'en multipliant le signal d'une oreille par l'HRTF de l'oreille opposée, mesurée à la direction de la source, on obtient :

$$\begin{aligned}\lambda(f,\hat{\theta},\hat{\phi}) &= S_L(f,\hat{\theta},\hat{\phi})\ HRTF_R(f,\hat{\theta},\hat{\phi}) = \\ &\quad S_o(f)\ HRTF_L(f,\hat{\theta},\hat{\phi})\ HRTF_R(f,\hat{\theta},\hat{\phi})\end{aligned} \tag{8}$$

et que

$$\begin{aligned}\chi(f,\hat{\theta},\hat{\phi}) &= S_R(f,\hat{\theta},\hat{\phi})\ HRTF_L(f,\hat{\theta},\hat{\phi}) = \\ &\quad S_o(f)\ HRTF_R(f,\hat{\theta},\hat{\phi})\ HRTF_L(f,\hat{\theta},\hat{\phi})\end{aligned} \tag{9}$$

alors

$$\lambda(f,\hat{\theta},\hat{\phi}) = \chi(f,\hat{\theta},\hat{\phi}) \tag{10}$$

Pour identifier la direction de la source, il est nécessaire d'effectuer le produit du signal avec l'ensemble des directions (θ,ϕ) pour lesquelles les HRTF sont connues et d'identifier la direction pour laquelle l'égalité (10) est vérifiée parmi cet ensemble.

Cette méthode permet d'obtenir une localisation avec des erreurs inférieures à 5° avec l'utilisation de 2 microphones et inférieurs à 2° avec 4 microphones pour un rapport signal à bruit proche de 40 dB.

4.2 Les Object-RTF dans la localisation des sources

Cette méthode part de l'hypothèse qu'il n'existe qu'une seule source à chaque instant par bande fréquentielle. Le traitement s'effectue alors sur des fenêtres temporelles dont la taille doit être déterminée en fonction de l'écart des capteurs et en fonction du nombre d'échantillons fréquentiels souhaité. Il est également possible de rajouter des zéros « zeroppadding » en fonction de la discrétisation spectrale souhaitée.

La relation (10) peut être généralisée à toute paire microphonique m, n en remplaçant les HRTF par les Object-RTF de la façon suivante :

$$\lambda(f, \hat{\theta}, \hat{\phi}) = S_m(f, \hat{\theta}, \hat{\phi})\ ObRTF_n(f, \theta', \phi') \quad (11a)$$

$$\chi(f, \hat{\theta}, \hat{\phi}) = S_n(f, \hat{\theta}, \hat{\phi})\ ObRTF_m(f, \theta', \phi') \quad (11b)$$

ou S_x sont les signaux issus d'un capteur x et $ObRTF_x(f, \theta', \phi')$ est la réponse en fréquence du capteur x pour la direction (f, θ', ϕ'). Comme précédemment la relation (10) est vérifiée

$$\text{ssi} \quad (\hat{\theta}, \hat{\phi}) = (\theta', \phi')$$

Ce cas théorique n'est pas toujours atteint. On cherche alors à trouver la meilleure estimation de la direction (θ', ϕ') parmi l'ensemble des directions (θ, ϕ) minimisant la distance δ entre ces directions.

Considérant $\beta_{\hat{n},m}$ comme le produit du signal S_n généré par une source placée à $(\hat{\theta}, \hat{\phi})$ avec l'ensemble des $ObRTF_m$ pour toutes les directions disponibles. C'est-à-dire :

$$\beta_{\hat{n},m} = S_n(f, \hat{\theta}, \hat{\phi})\ ObRTF_m(f, \theta, \phi), \quad (12)$$

où $\beta_{\hat{n},m}$ est représenté par un vecteur de longueur L correspondant au nombre des directions des Object-RTF.

On définit alors une distance $\delta(n, m)$ entre chaque paire d'éléments $\beta_{\hat{n},m}(l)$ et $\beta_{\hat{m},n}(l)$

$$\delta(n, m) = \delta\left(\beta_{\hat{n},m}, \beta_{\hat{m},n}\right) \quad (13)$$

On considère alors que la position de la source $(\hat{\theta}, \hat{\phi})$ est la direction où $\delta(n, m)$ est minimale.

Lorsque plus de deux capteurs sont utilisés simultanément, l'analyse s'effectue sur l'ensemble des paires microphoniques. La direction retenue est celle pour laquelle la valeur $\delta(m, n)$ est la plus faible.

4.3 Définition du critère de distance δ

Les données manipulées étant des valeurs complexes, la norme des vecteurs n'est pas le meilleur indicateur de ressemblance de ces derniers, trois indicateurs de distance ont été alors comparés.

Distance angulaire : Dans un premier temps, nous avons proposé l'utilisation du cosinus de l'angle entre les deux valeurs complexes comme indicateur de distance δ entre ces deux valeurs. Celui-ci est calculé grâce au produit scalaire.

$$\delta(A, B) = \cos(\widehat{AB}) = \frac{A \cdot B}{|A||B|} \quad (14)$$

avec

$$A = S_m(f, \hat{\theta}, \hat{\phi}) ObRTF_n(f, \theta, \phi)$$
$$B = S_n(f, \hat{\theta}, \hat{\phi}) ObRTF_m(f, \theta, \phi)$$

Cette relation présente l'avantage de fournir un indice normalisé de ressemblance qui vaut 0 lorsque les deux vecteurs sont colinéaires et qui vaut 1 lorsqu'ils sont perpendiculaires. Néanmoins, cette méthode ne rend pas compte de la norme des vecteurs.

Tanimoto : Afin de rassembler la magnitude et l'angle des vecteurs dans un seul indicateur. Il est proposé d'utiliser un deuxième indicateur de distance $D_T(A, B)$ [2] issu de l'indice de Tanimoto $T(A, B)$ qui est défini par :

$$D_T(A, B) = 1 - T(A, B) \quad (15)$$

$$= 1 - \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}, \quad (16)$$

où $\cdot$ définit le produit scalaire et $|\ |$ la norme du vecteur.

Quotient : Une dernière mesure de distance, appelée ici la méthode du quotient, est proposée. Partant de l'hypothèse définie dans l'équation (10) avec les paramétrés de (11) :

$$\frac{S_n(f, \hat{\theta}, \hat{\phi})\ ObRTF_m(f, \hat{\theta}, \hat{\phi})}{S_m(f, \hat{\theta}, \hat{\phi})\ ObRTF_n(f, \hat{\theta}, \hat{\phi})} = 1 \quad (17)$$

la distance $\delta\left(\beta_{\hat{n},m}, \beta_{\hat{m},n}\right)$ est alors définie par :

$$\delta = \frac{S_n(f, \hat{\theta}, \hat{\phi})\ ObRTF_m(f, \hat{\theta}, \hat{\phi})}{S_m(f, \hat{\theta}, \hat{\phi})\ ObRTF_n(f, \hat{\theta}, \hat{\phi})} - 1 \quad (18)$$

permettant de ne calculer qu'une seule fois les rapports $\frac{ObRTF_n(f,\theta,\phi)}{ObRTF_m(f,\theta,\phi)}$ car ils sont constants tout au long de l'analyse et sont indépendants des signaux S_n. Ce critère permet de réduire considérablement le nombre d'opérations en allégeant ainsi la complexité.

5 Evaluation des performances des Object-RTF pour la localisation des sources

Des simulations numériques ont été mises en place pour effectuer les évaluations. Le signal test est une source large bande tournant autour du microphone à des élévations variant par palier de 20° partant de l'hémisphère sud vers le zénith.

Nous étudions notamment ici, le meilleur critère permettant d'évaluer la distance δ (cf. équation (13))

5.1 Critères d'évaluation

Les critères d'évaluations introduits dans [11] sont utilisés. E_t est la distance angulaire totale sur la sphère par rapport à la position réelle de la source et E_{75} est l'erreur E_t obtenue pour au moins 75 % du spectre calculé en tiers d'octaves.

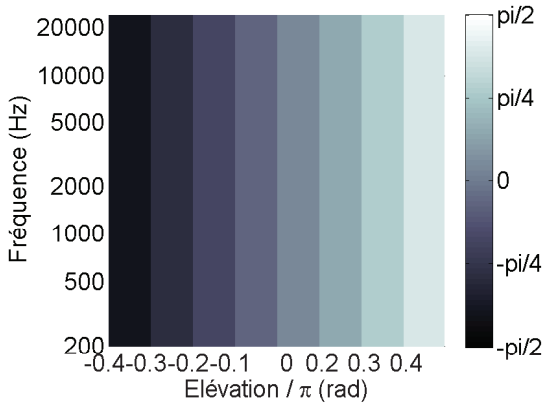


FIGURE 2 – Élévation calculée, δ évalué avec le coefficient de Tanimoto

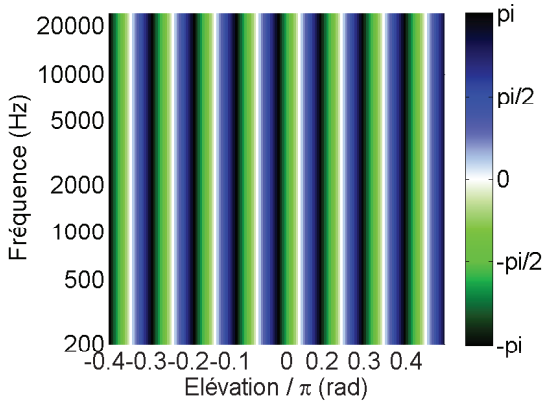


FIGURE 3 – Azimuth calculé, δ évalué avec le coefficient de Tanimoto

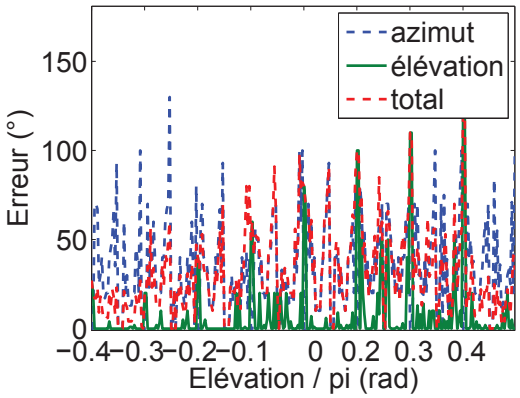


FIGURE 4 – Erreur de localisation E_{75}, δ calculé avec le critère du quotient

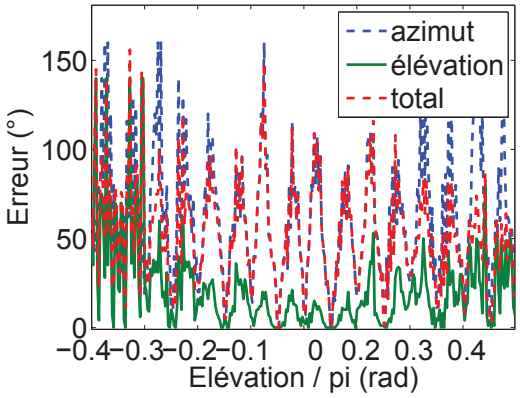


FIGURE 5 – Erreur de localisation E_{75}, résultats obtenus avec un RSB de 30 dB pour une source perturbatrice placée à $\theta = 0$ et $\phi = 0$

5.2 Performances

Les critères de distance définis dans le paragraphe 4.3 ont été évalués. L'utilisation des critères de Tanimoto et du quotient permettent une localisation correcte et sans ambiguïté dans une configuration idéale (cf. figures 2 et 3). L'utilisation de la distance angulaire laisse apparaître des nombreuses erreurs de localisation (cf. figure 4) où l'E_{75} moyen est de 31°. Ces erreurs apparaissent notamment lorsque la source s'éloigne de l'azimut de 0° et elles augmentent avec l'élévation pour atteindre des erreurs maximums de localisation E_{75} de 130°.

Ces résultats nous permettent d'écarter l'utilisation du critère défini par la distance angulaire, nous permettant de limiter les études ultérieures aux coefficients de Tanimoto et au critère du quotient.

Rapport signal à bruit : Les résultats de localisation obtenus avec les coefficients retenus (ie. Tanimoto et Quotient) sont identiques quelque soit le rapport signal à bruit. L'erreur E_{75} moyennée sur l'ensemble des directions est de 64,5° lorsque le rapport signal à bruit (RSB) est de 20 dB et atteint 10,0° avec un RSB de 30 dB. Cette erreur est nulle quand la source sonore se trouve au même azimut de la source perturbatrice et maximale lorsqu'elles sont diamétralement opposées.

Si uniquement les deux microphones placés sur le plan horizontal sont utilisés, l'indicateur E_{75} moyen atteint 64,5° pour un RSB de 20 dB et de 15° pour un RSB de 30 dB. Lorsque le RSB est de 40 dB le E_{75} moyen est de 0,5° et l'erreur maximum est de 26° confortant ainsi les résultats de Keyruz et McDonald [6, 8].

Complexité : L'utilisation de la méthode du quotient permet le calcul de la matrice $\frac{ObRTF_n(f,\theta,\phi)}{ObRTF_m(f,\theta,\phi)}$ (cf. paragraphe 4.3) ce qui permet de gagner considérablement en complexité algorithmique. En effet, les calculs réalisés avec cette méthode sont effectués dans environ un tiers du temps utilisé par la méthode de Tanimoto.

6 Conclusion

Afin de proposer une nouvelle méthode permettant la captation du son spatialisé avec des terminaux mobiles, nous avons développé une méthode permettant l'encodage de la scène sonore avec un signal audio représentatif de la scène et des paramètres de localisation des sources qui la composent.

Un dispositif microphonique composé de 3 capsules a été proposé auparavant ainsi que les algorithmes de localisation associés. Ces derniers étant très

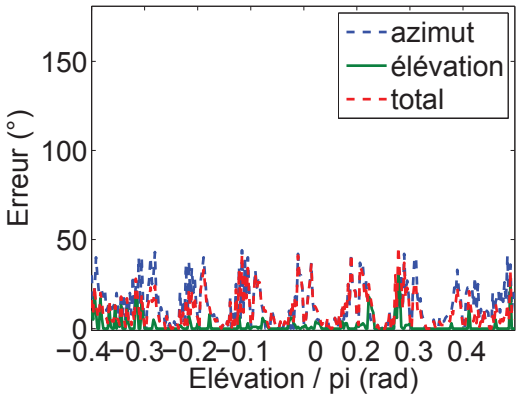


FIGURE 6 – Erreur de localisation E_{75}, résultats obtenus avec un RSB de 30 dB pour une source perturbatrice placée à $\theta = 0$ et $\phi = 0$

contraignants, une nouvelle méthode appelée méthode des Objets-RTF a été proposée ici. Elle permettant de prendre en compte les caractéristiques réelles des capteurs. Celle-ci est une généralisation de la méthode de HRTF de Keyruz et McDonald.

La direction de la source est obtenue en cherchant la similitude parmi l'ensemble de object-RTF par un critère de distance. Nous avons proposé et évalué ici trois critères permettant la mesure de cette distance (ie. Distance Angulaire, Tanimoto et Quotient). Dans des condition idéales les critères de Tanimoto et du quotient donnent des résultats parfaits. Dans les mêmes conditions, le troisième critère donne des résultats comportant une erreur moyenne de 31° ce qui permet de l'écarter. La robustesse de la méthode été évalué en présence d'une source perturbatrice avec les deux critères de distance retenus. les résultats sont identiques pour les deux critères quel que soit le rapport signal à bruit. Les erreurs moyennes de localisations rencontrées sont inférieures à 16° pour un RSB de 30 dB avec l'utilisation de 2 capteurs et de 10° lorsque 3 capteurs sont utilisés.

En termes de complexité, la méthode du quotient est à privilégier car elle permet une réduction considérable du temps de calcul par rapport à la méthode de Tanimoto.

Des études complémentaires doivent être effectuées afin d'évaluer les performances de cette méthode avec des dispositifs réels ce qui implique la mesure des Object-RTF du dispositif microphonique.

Références

[1] S. Berge and N. Barrett. A new method for b-format to binaural transcoding. In *AES 40th international conference*, Tokio, Japan, Oct. 2010.

[2] R. Camacho and al. Assessing the effect of 2D fingerprint filtering on ILP-Based structure-activity relationships toxicity studies in drug design. In *5th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2011)*, volume 93 of *Advances in Intelligent and Soft Computing*, pages 355–363. Springer Berlin Heidelberg, Jan. 2011.

[3] M. A. Gerzon. The design of precisely coincident microphone arrays for stereo and surround sound. *Audio Engineering Society Convention 50*, 50 :50, 1975.

[4] M. A. Gerzon and P. G. Craven. Coincident microphone simulation covering three dimensional space and yelding various directionall outputs, Aug. 1977.

[5] F. Keyrouz and K. Diepold. An enhanced binaural 3D sound localization algorithm. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, page 662–665, 2006.

[6] F. Keyrouz, K. Diepold, and P. Dewilde. Robust 3d robotic sound localization using state-space hrtf inversion. In *Robotics and Biomimetics, 2006. ROBIO'06. IEEE International Conference on*, page 245–250, 2006.

[7] E. H. Langendijk and A. W. Bronkhorst. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *The Journal of the Acoustical Society of America*, 107 :528, 2000.

[8] J. A. MacDonald. A localization algorithm based on head-related transfer functions. *The Journal of the Acoustical Society of America*, 123(6) :4290, 2008.

[9] P. Majdak, P. Balazs, and B. Laback. Multiple exponential sweep method for fast measurement of head related transfer functions. *Journal Audio Engeneering Society*, 2007.

[10] J. Palacino and R. Nicol. Full 3D sound pick-up with a small microphone array : Prototype outline and preliminary assessment. In *Proceedings of the International Conference on Acoustics*, Merano, Mar. 2013.

[11] J. Palacino and R. Nicol. Spatial sound pick-up with a low number of microphones. In *Proceedings of Meetings on Acoustics*, volume 19, page 055078, Montreal, June 2013.

[12] J. Palacino, R. Nicol, M. Emerit, and L. Gros. Perceptual assessment of binaural decoding of first-order ambisonics. In *Acoustics 2012*, Nantes, France, Apr. 2012.

[13] J.-M. PERNAUX. *Spatialisation du son par les techniques binaurales : application aux services de télécommunications.* PhD, I.N.P.G, Grenoble, May 2003.

[14] V. Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. In *Proc. of the AES 28th Int. Conf, Pitea, Sweden*, 2006.

E.6 Brevet : Acquisition de données sonores spatialisées [Palacino and Nicol, 2012]

R E P U B L I Q U E F R A N C A I S E

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITE

Réception électronique de la soumission

Il est certifié par la présente qu'une demande de brevet (ou d'un certificat d'utilité) a été reçue par le biais du dépôt électronique sécurisé de l'INPI. Après réception, un numéro d'enregistrement et une date de réception ont été automatiquement attribués.

Numéro de demande	1260898	
Numéro de soumission	1000170765	
Date de réception	16 novembre 2012	
Vos références	BFF120477/TM	
Demandeur	FRANCE TELECOM	
Pays	FR	
Titre de l'invention	ACQUISITION DE DONNÉES SONORES SPATIALISÉES	
Documents envoyés	package-data.xml	requetefr.xml
	application-body.xml	fr-fee-sheet.xml
	requetefr.pdf (3 p.)	validation-log.xml
	comment.pdf (1 p.)	indication-bio-deposit.xml
	design.pdf (1 p.)	fr-office-specific-info.xml
	dessins.pdf (4 p.)	textebrevet.pdf (41 p.)
Déposé par	EMAIL=info@plass.com,CN=Albert HASSINE,O=CABINET PLASSERAUD,C=FR	
Méthode de dépôt	Dépôt électronique	
Date et heure de réception électronique	16 novembre 2012, 10:18:57 (CET)	
Empreinte officielle du dépôt	BF:5D:96:33:65:CE:8F:45:9B:BC:39:12:3B:76:EB:59:DC:98:D7:14	

/INPI, section dépôt/

INPI - 26 BIS, RUE DE SAINT PETERSBOURG 75800 PARIS CEDEX 08 - TELEPHONE : 0 820 213 213 - TELECOPIE : 01 53 04 52 65
ETABLISSEMENT PUBLIC NATIONAL CREE PAR LA LOI No 51-444 DU 19 AVRIL 1951

INSTITUT NATIONAL
DE LA PROPRIÉTÉ
INDUSTRIELLE

BREVET D'INVENTION

Code de la propriété intellectuelle - Livre VI

cerfa

15 rue des Minimes - CS 50001 - 92677 Courbevoie Cedex

Pour vous informer : INPI Direct 0820 210 211

Pour déposer par télécopie : 33 (0)1 56 65 86 00

REQUÊTE EN DÉLIVRANCE

REMISE DES PIÈCES DATE N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI	**1NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE** Cabinet Plasseraud 52 rue de la Victoire 75440 PARIS CEDEX 09 FR
Vos références pour ce dossier BFF120477/TM	

2 NATURE DE LA DEMANDE	
Nature	Brevet d'invention
3 TITRE DE L'INVENTION	
Titre	ACQUISITION DE DONNÉES SONORES SPATIALISÉES
4 PRIORITÉ	
5-1 DEMANDEUR	**Personne morale**
Nom	FRANCE TELECOM
Affaire suivie par	
Rue	78 Rue Olivier de Serres
Code postal et ville	75015 PARIS
Pays	FR
Nationalité	FR
Forme juridique	Société anonyme
N° SIREN	380 129 866
Code APE-NAF	
N° de téléphone	
N° de télécopie	
Courrier électronique	
6 MANDATAIRE	
Nom	Cabinet Plasseraud
Qualité	Cabinet CPI : 04-0603, pas de pouvoir
Rue	52 rue de la Victoire

Code postal et ville	75440 PARIS CEDEX 09			
Pays	FR			
N° de téléphone	00 33 1 40 16 70 00			
N° de télécopie	00 33 1 42 80 01 59			
Courrier électronique	info@plass.com			
7 RAPPORT DE RECHERCHE				
Type d'établissement	Établissement immédiat			
8 RÉDUCTION DU TAUX DES REDEVANCES				
9 DÉPÔT DE MATIÈRE BIOLOGIQUE				
10 SÉQUENCES DE NUCLÉOTIDES ET/OU D'ACIDES AMINÉS				
11 DOCUMENTS ET FICHIERS JOINTS	**Fichier électronique**		**Détails**	
Inventeur	Design.PDF			
Fichier corps du texte	textebrevet.pdf		page(s) 41, D 32, R 8, AB 1	
Drawings	dessins.pdf		page(s) 4, Abrégé : page 1, Fig. 1	
12 MODE DE PAIEMENT				
Mode de paiement	Prélèvement du compte client			
Numéro du compte	3200			
13 REDEVANCES JOINTES	Devise	Taux	Quantité	Montant à payer
062 Dépôt d'une demande électronique	EURO	26	1	26
063 Rapport de recherche	EURO	500	1	500
068 Revendication à partir de la 11ème	EURO	40	5	200
Total	EURO			726
14 ANNOTATION				
15 DATE ET SIGNATURE				
Signé numériquement par	Subject: FR, CABINET PLASSERAUD, Albert HASSINE; Issuer: FR, INPI, INPI-EN-LIGNE 1.1			
Date	16 November 2012			
Signataire	Mandataire			

pour identifier le titulaire de la demande et son éventuel mandataire.

INSTITUT NATIONAL
DE LA PROPRIÉTÉ
INDUSTRIELLE

15 rue des Minimes - CS 50001 - 92677 Courbevoie Cedex

Pour vous informer : INPI Direct 0820 210 211

Pour déposer par télécopie : 33 (0)1 56 65 86 00

BREVET D'INVENTION

Code de la propriété intellectuelle - Livre VI

cerfa

DÉSIGNATION D'INVENTEUR(S)

Vos références pour ce dossier	BFF120477/TM
N° D'ENREGISTREMENT NATIONAL	
TITRE DE L'INVENTION	
Titre	ACQUISITION DE DONNÉES SONORES SPATIALISÉES
LE(S) DEMANDEUR(S)	FRANCE TELECOM
DÉSIGNE(NT) EN TANT QU'INVENTEUR(S)	
INVENTEUR 1	
Nom	PALACINO
Prénom	Julian
Rue	52, rue de Morlaix
Code postal et ville	22310 PLESTIN LES GREVES
Pays	FR
INVENTEUR 2	
Nom	NICOL
Prénom	Rozenn
Rue	2, place du Martray
Code postal et ville	22450 LA ROCHE DERRIEN
Pays	FR
DATE ET SIGNATURE	
Signé numériquement par	Subject: FR, CABINET PLASSERAUD, Albert HASSINE; Issuer: FR, INPI, INPI-EN-LIGNE 1.1
Date	16 November 2012
Signataire	Mandataire

Conformément aux dispositions de la loi n° 78-17 du 6.01.1978 modifiée relative à l'informatique, aux fichiers et aux libertés, vous bénéficiez d'un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI. Les données à caractère personnel que vous êtes tenu(e) de nous fournir dans ce formulaire sont exclusivement utilisées pour identifier le titulaire de la demande et son éventuel mandataire.

265

INSTITUT NATIONAL
DE LA PROPRIÉTÉ
INDUSTRIELLE

15 rue des Minimes - CS 50001 - 92677 Courbevoie Cedex

Pour vous informer : INPI Direct 0820 210 211

Pour déposer par télécopie : 33 (0)1 56 65 86 00

BREVET D'INVENTION

Code de la propriété intellectuelle - Livre VI

cerfa

JOURNAL DE VALIDATIONS

1 REQUÊTE	
2 NOMS	
VERT	FRANCE TELECOM: L'indication d'une adresse de messagerie électronique est recommandée.
VERT	FRANCE TELECOM: L'indication d'un numéro de fax est recommandée.
VERT	FRANCE TELECOM: L'indication d'un numéro de téléphone est recommandée.
VERT	FRANCE TELECOM: Le code APE-NAF n'a pas été renseigné.
VERT	Julian PALACINO: L'indication d'une adresse de messagerie électronique est recommandée.
VERT	Julian PALACINO: L'indication d'un numéro de fax est recommandée.
VERT	Julian PALACINO: L'indication d'un numéro de téléphone est recommandée.
VERT	Rozenn NICOL: L'indication d'une adresse de messagerie électronique est recommandée.
VERT	Rozenn NICOL: L'indication d'un numéro de fax est recommandée.
VERT	Rozenn NICOL: L'indication d'un numéro de téléphone est recommandée.
3 PRIORITÉ	
4 MATIÈRE BIOLOGIQUE	
5 DOCUMENTS	
6 PAIEMENT DES TAXES	
7 ANNOTATIONS	

Acquisition de données sonores spatialisées

La présente invention concerne le domaine des technologies de prise de son et des technologies de traitement audio associées.

Elle concerne en particulier, mais non exclusivement, un procédé de traitement de données sonores issues d'une scène sonore tridimensionnelle capable d'extraire une information de position spatiale de sources sonores. Elle trouve des applications aussi bien pour la prise de son spatialisée dans le cadre de services conversationnels, que pour l'enregistrement de contenus audio 3D (par exemple concert, paysage sonore, etc).

Différentes méthodes de prise de son spatialisé sont connues. Certaines cherchent à saisir les informations exploitées par le système auditif (technologie binaurale par exemple) tandis que d'autres décomposent le champ acoustique de façon à restituer une information spatiale plus ou moins riche qui sera interprétée par l'auditeur (technologie ambisonique par exemple).

Une première méthode consiste en une prise de son stéréophonique. Les différences de phase et/ou de temps, et d'amplitude entre des signaux issus de deux microphones sont exploitées afin de recréer des stimuli constituant une approximation grossière de l'écoute naturelle. Ces signaux sont restitués sur une paire de haut-parleurs toujours placés face à l'auditeur et alignés sur le plan horizontal. Dans une telle configuration, toute information provenant de l'arrière de l'auditeur et toute notion d'élévation sont perdues. Afin d'enrichir l'arrière de la scène sonore, de nombreuses solutions ont été proposées. En particulier, de telles solutions consistent généralement en une augmentation du nombre de capteurs visant les directions recherchées. On peut également prévoir un matriçage des signaux stéréophoniques afin d'enrichir l'arrière de la scène sonore. De telles solutions ont donné naissance aux systèmes quadriphoniques, 5.1 et 7.1.

Cependant, la prise de son stéréophonique est toujours limitée au plan horizontal frontal, ou horizontal dans le cas des extensions multicanaux de types 5.1. En d'autres termes, dans le meilleur des cas, en coordonnées sphériques, elle n'est capable d'identifier que l'information d'azimut des sources

sonores (les coordonnées des sources dans un plan horizontal x-y), sans pour autant pouvoir identifier leur information d'élévation.

Une deuxième méthode consiste en une prise de son binaurale. La technologie binaurale permet une captation et une restitution imitant une écoute naturelle, permettant notamment la localisation d'une source dans tout l'espace entourant l'auditeur, en utilisant uniquement deux microphones. Les microphones sont placés dans les oreilles d'une personne ou d'un mannequin afin d'enregistrer la scène acoustique et les indices acoustiques de la localisation naturelle.

La prise de son directe utilisant la technologie binaurale présente cependant différents inconvénients. En effet, lorsque la prise de son est effectuée sur une tête naturelle, la personne portant les microphones doit rester immobile, contrôler sa respiration et éviter de déglutir afin de ne pas détériorer la qualité de l'enregistrement. L'utilisation d'une tête artificielle est difficilement envisageable lorsque l'on recherche une utilisation discrète et portative. Au moment de la restitution, l'incompatibilité des fonctions de transferts relatives à la tête de l'auditeur (« *Head Related Transfer Function* » en anglais ou HRTF) entre le dispositif de captation et l'auditeur final tend à fausser la localisation des sources. D'autre part, lorsque l'auditeur final bouge la tête, l'ensemble de la scène sonore se déplace.

Ainsi, bien que la prise de son binaurale soit capable d'encoder l'information spatiale des sources dans tout l'espace tridimensionnel, un tel encodage est spécifique de la morphologie de la personne ou du mannequin qui a servi à l'enregistrement. Aucune solution satisfaisante n'a été proposée à ce jour afin de remédier à ces limitations. Un inconvénient supplémentaire est que l'enregistrement binaural ne peut être écouté que sur un équipement spécifique dédié tel qu'un casque ou un système de haut-parleurs associés à un prétraitement.

Une troisième méthode consiste en une prise de son ambisonique par captation du champ acoustique. Une telle technologie fut introduite dans le document US 4,042,779 pour les harmoniques sphériques du premier ordre, et son extension aux ordres supérieurs HOA fut décrite par exemple dans le document J. Daniel, « *Représentation de champs acoustiques, application à la*

transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia », Université Paris 6, Paris, 2001. Ces documents permettent une acquisition plus ou moins précise de la scène sonore en fonction de l'ordre des harmoniques sphériques utilisées.

Cependant, une telle technologie présente l'inconvénient d'utiliser un grand nombre de capteurs, qui est fonction de l'ordre désiré. L'utilisation de la technologie ambisonique à l'ordre 1 a été largement exploitée en raison du faible nombre de capteurs requis (quatre microphones, voir US 4,042,779) pour sa mise en œuvre. Des signaux issus des quatre microphones sont dérivés par matriçage (encodage), les quatre signaux définissant le format B de la technologie ambisonique. Les signaux dérivés par matriçage correspondent aux signaux qui auraient été enregistrés par un microphone omnidirectionnel et trois microphones à gradient de vitesse orientés selon les axes x, y et z. Les quatre signaux dérivés sont enregistrés et peuvent ensuite être restitués à un auditeur en utilisant un système de haut-parleurs distribués de façon arbitraire grâce à une matrice de décodage. Les haut-parleurs ainsi choisis peuvent également être obtenus sous la forme de sources virtuelles pour une restitution binaurale utilisant les fonctions de transfert HRTF relatives à la position de chaque source.

Ainsi, la prise de son ambisonique est aussi capable d'encoder l'information spatiale des sources dans tout l'espace 3D mais elle présente l'inconvénient de nécessiter un nombre important de capteurs, à savoir au minimum 4, et potentiellement un nombre encore plus important lorsqu'une précision spatiale satisfaisante est recherchée.

Il peut également être envisagé des post-traitements associés à la prise de son spatialisé, afin de remédier aux inconvénients détaillés ci-dessus.

En particulier, de tels traitements sont appliqués afin d'améliorer l'extraction de l'information spatiale. Jusqu'à présent, des post-traitements ont été appliqués à des signaux de type ambisonique, car ces derniers donnent accès à une représentation physique des ondes acoustiques.

Le document de V. Pulkki, « *Directional audio coding in spatial sound reproduction and stereo upmixing* », in Proc. of the AES 28th Int. Conf, Pitea, Sweden, 2006, propose une méthode pour extraire des signaux du format B les

informations de localisation des sources. L'objectif d'une telle méthode est l'obtention d'une représentation plus compacte de la scène sonore tridimensionnelle (compression de l'information), dans laquelle les quatre signaux issus du format B sont ramenés à un unique signal monophonique accompagné d'un signal comportant des informations de localisation des sources sonores.

Un perfectionnement de cette méthode a été proposé dans le document de N. Barrett et S. Berge, « *A new method for B-format to binaural transcoding* », in 40th AES International conference. Tokyo, Japan, 2010, p. 8–10. Ce perfectionnement prévoit d'utiliser les informations de localisation pour spatialiser les sources sonores virtuelles en vue d'une restitution sur haut-parleurs ou transcodage binaural. Les sources sonores virtuelles sont ainsi re-spatialisées a posteriori conformément à leur position identifiée, dans le format de spatialisation associé au dispositif de restitution.

Toutefois, que ce soit la méthode précédente ou sa version perfectionnée, la position des sources est déterminée avec une ambiguïté (typiquement une ambiguïté angulaire de ±π/2 sur l'angle d'azimut dans le document de V. Pulkki), qui n'est pas résolue. La position de la source sonore n'est alors pas connue avec certitude.

La présente invention vient améliorer la situation.

Un premier aspect de l'invention concerne un procédé de traitement de données pour la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le procédé comprenant les étapes suivantes :

- obtenir au moins un premier signal et un deuxième signal à partir du signal sonore capté selon des directivités différentes par un premier capteur et un deuxième capteur ;
- déduire des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, l'expression comportant une incertitude sur la coordonnée spatiale ;
- déterminer une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les

premier et deuxième capteurs ;

- déterminer la première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

Ainsi, la présente invention prévoit, à partir d'un nombre réduit de capteurs (au moins deux) de lever une incertitude sur une expression d'une coordonnée spatiale d'une source sonore, et ce, par la détermination d'une information supplémentaire qui exploite les caractéristiques des signaux reçus respectivement sur les capteurs. Par exemple, l'incertitude peut être due à une fonction cosinus inverse comme c'est le cas dans le document de V. Pulkki.

La présente invention permet ainsi d'améliorer la précision lors de la localisation de la source (détermination d'une coordonnée spatiale de la source). De plus, la présente invention est adaptable à toute unité microphonique comprenant les premier et deuxième capteurs. Ainsi, comme détaillé dans ce qui suit, les premier et deuxième capteurs peuvent être des microphones cardioïdes, des microphones bidirectionnels, ou être intégrés dans un microphone ambisonique d'ordre 1 ou d'ordre supérieur à 1 plus généralement.

Selon un mode de réalisation, l'espace étant orienté selon trois axes x, y et z, les premier et deuxième capteurs sont des microphones cardioïdes, le premier microphone cardioïde étant situé à une première position de l'espace et orienté selon l'axe x dans un sens croissant et le deuxième microphone cardioïde étant situé à une deuxième position de l'espace et orienté selon l'axe x dans un sens décroissant, le procédé peut comprendre initialement :

- modifier la première ou la deuxième position en vue d'introduire un décalage selon l'axe y entre le premier microphone cardioïde et le deuxième microphone cardioïde.

Les premier et deuxième signaux correspondent aux signaux sonores captés respectivement par les premier et deuxième microphones cardioïdes et l'information supplémentaire est le signe d'une différence entre des phases respectives des premiers et seconds signaux.

En décalant les microphones cardioïdes selon un axe perpendiculaire à l'axe d'orientation des microphones cardioïdes, l'invention permet l'introduction d'un retard entre les signaux captés par ces microphones, retard qui constitue

une information supplémentaire à partir de laquelle il est possible de déterminer avec certitude la première coordonnée spatiale.

Des première et deuxième coordonnées spatiales de la source sonore peuvent être respectivement les coordonnées sphériques d'azimut θ et d'élévation ϕ. Comme détaillé dans ce qui suit, les coordonnées spatiales peuvent être exprimées dans le domaine temporel ou dans le domaine fréquentiel. En complément, un troisième signal s_{card3} capté par un troisième microphone cardioïde orienté selon l'axe z dans un sens croissant peut être obtenu. Le premier signal capté par le premier microphone cardioïde étant noté s_{card1} et le second signal capté par le second microphone cardioïde étant noté s_{card2}, les expressions des coordonnées sphériques d'azimut θ et d'élévation ϕ peuvent être données par :

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi)$$

$$\phi = \sin^{-1}[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1]$$

$$\theta = \cos^{-1}[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

Le signe de la coordonnée sphérique d'azimut θ est donné par la différence de phases entre les premier et deuxième signaux.

L'expression de la coordonnée sphérique d'azimut θ présente une ambiguïté résultant de la fonction cosinus inverse. Cette ambiguïté est levée par exemple en exploitant la différence de phases entre les premier et deuxième signaux qui donne le signe de la coordonnée sphérique d'azimut θ. Toutefois, l'invention n'est aucunement restreinte à la prise en compte de la différence de phase entre les premier et deuxième signaux : elle s'applique à toute information supplémentaire permettant d'identifier le demi-espace dans lequel est située la source sonore, ce qui permet de lever l'ambiguïté précitée.

Ainsi, la présente invention permet de déterminer complètement la direction de la source sonore (connaissance des coordonnées sphériques d'azimut θ et d'élévation ϕ) à partir de seulement trois capteurs, à savoir les trois microphones cardioïdes, tout en levant l'incertitude sur la coordonnée sphérique d'azimut θ. A noter qu'aucune restriction n'est attachée aux

coordonnées considérées : la présente invention est applicable en coordonnées cylindriques ou cartésiennes.

En variante, des première et deuxième coordonnées spatiales de la source sonore peuvent être respectivement les coordonnées sphériques d'azimut θ et d'élévation ϕ, le premier signal capté par le premier microphone cardioïde étant noté s_{card1} et le second signal capté par le second microphone cardioïde étant noté s_{card2}, l'expression de la coordonnée sphérique d'azimut θ est donnée par :

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi)$$

$$\theta = \cos^{-1}\left[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}\right]$$

La coordonnée sphérique d'élévation ϕ peut être fixée arbitrairement et le signe de la coordonnée sphérique d'azimut θ peut être donné par la différence de phases entre les premier et deuxième signaux.

Cette variante permet de diminuer le nombre de capteurs à uniquement deux microphones cardioïdes, au détriment de la précision liée à la coordonnée sphérique d'élévation ϕ, tout en levant l'incertitude sur la coordonnée sphérique d'azimut θ.

Selon un mode de réalisation de l'invention, l'espace étant orienté selon trois axes x, y et z, des première et deuxième coordonnées spatiales de la source sonore peuvent être les coordonnées sphériques d'azimut θ et d'élévation ϕ, les premier et deuxième capteurs peuvent être des capteurs bidirectionnels, le premier capteur étant orienté selon l'axe x et captant le signal noté s_{bi1} et le deuxième capteur étant orienté selon l'axe y et captant le signal noté s_{bi2}.

Un troisième capteur cardioïde peut être dirigé selon l'axe z croissant et être apte à capter un signal noté s_{card3}. Les premiers et second signaux peuvent être notés respectivement $s_{cardvirt1}$ et $s_{cardvirt2}$ et être obtenus par:

$$s_{cardvirt1}(t,\theta,\phi) = \frac{s_0(t)}{2}\left(1 + \frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}\right) ;$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{s_0(t)}{2}\left(1 - \frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}\right) ;$$

avec $s_0(t) = \frac{s_{bi1}^2(t,\theta,\phi) + s_{bi2}^2(t,\theta,\phi) + 4s_{card3}^2(t,\theta,\phi)}{4s_{card3}^2(t,\theta,\phi)}$.

Les expressions des coordonnées sphériques d'azimut θ et d'élévation ϕ peuvent être données par:

$$\phi = \sin^{-1}[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1]$$

$$\theta = \cos^{-1}[\frac{s_{cardvirt1}^2(t,\theta,\phi) - s_{cardvirt2}^2(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

L'information supplémentaire pour lever l'ambiguïté peut être déterminée à partir d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux s_{bi1} et s_{bi2}.

Ainsi, la présente invention est applicable lorsque le signal sonore est initialement capté par des microphones bidirectionnels. En effet, en obtenant les premier et deuxième signaux, des microphones cardioïdes virtuels sont synthétisés, ce qui permet de revenir à des expressions semblables au premier mode de réalisation pour les coordonnées sphériques d'azimut et d'élévation. En revanche, dans le cas de microphones bidirectionnels, il n'est pas possible d'introduire un retard, et l'invention prévoit alors la prise en compte d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux captés par ces microphones, afin de lever l'incertitude sur la détermination de la coordonnée sphérique d'azimut. A nouveau, trois capteurs seulement permettent une détermination complète de la direction de la source sonore.

En variante, l'espace étant orienté selon trois axes x, y et z, les première et deuxième coordonnées spatiales peuvent être les coordonnées sphériques d'azimut θ et d'élévation ϕ, les premier et deuxième capteurs peuvent être des capteurs bidirectionnels, le premier capteur étant orienté selon l'axe x et captant le signal noté s_{bi1} et le deuxième capteur étant orienté selon l'axe y et captant le signal noté s_{bi2}, les premiers et second signaux peuvent être notés respectivement $s_{cardvirt1}$ et $s_{cardvirt2}$ et peuvent être obtenus de la manière suivante :

$$s_{cardvirt1}(t,\theta,\phi) = \frac{s_0(t)}{2}(1+\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \ ;$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{s_0(t)}{2}(1-\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \ ;$$

avec $s_0(t) = \sqrt{s_{bi1}^2(t,\theta,\phi) + s_{bi2}^2(t,\theta,\phi)}$.

L'expression de la coordonnée sphérique d'azimut θ peut être donnée par:

$$\theta = \cos^{-1}[\frac{s_{cardvirt1}^2(t,\theta,\phi) - s_{cardvirt2}^2(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

La coordonnée sphérique d'élévation ϕ est fixée arbitrairement et l'information supplémentaire peut être déterminée à partir d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux s_{bi1} et s_{bi2}.

Ainsi, il est possible de diminuer le nombre de capteurs à deux microphones bidirectionnels tout en levant l'incertitude sur la coordonnée sphérique d'azimut θ, au détriment de la précision quant à la détermination de la coordonnée sphérique d'élévation ϕ.

En complément, l'information supplémentaire peut être une deuxième expression de la coordonnée sphérique d'azimut θ :

$$\theta(\omega) = \tan^{-1}(\frac{I_y(\omega)}{I_x(\omega)})$$

ω étant une pulsation du signal sonore émis par la source,

$I_y(\omega)$ étant la composante selon la coordonnée y du vecteur d'intensité acoustique du signal sonore, donnée par :

$$I_y(\omega) = \frac{1}{2\rho c}\mathrm{Re}[S_0^*(\omega)S_{bi2}(\omega)] \ ;$$

$I_x(\omega)$ étant la composante selon la coordonnée x du vecteur d'intensité acoustique du signal sonore, donnée par :

$$I_x(\omega) = \frac{1}{2\rho c}\mathrm{Re}[S_0^*(\omega)S_{bi1}(\omega)] \ ;$$

$S_0(\omega)$, $S_{bi1}(\omega)$ et $S_{bi2}(\omega)$ désignant les transformées de Fourier des signaux $s_0(t)$, $s_{bi}1(t)$ et $s_{bi2}(t)$ respectivement.

Ainsi, en obtenant une expression supplémentaire sur la coordonnée sphérique d'azimut θ, il est possible de lever l'incertitude liée à la première expression comprenant la fonction cosinus inverse. En effet, bien que la fonction tangente inverse présente également une incertitude, la fonction tangente inverse et la fonction cosinus inverse permettent d'obtenir deux estimations de la coordonnée sphérique d'azimut θ qui sont complémentaires. Par exemple, comme détaillé ultérieurement, l'utilisation d'un tableau permet de différencier quatre cas de figures, selon les intervalles dans lesquels se situent les deux estimations de la coordonnée sphérique d'azimut θ. Une valeur désambiguïsée de la coordonnée sphérique d'azimut θ peut être déterminée. En complément, il est possible de prendre en compte des facteurs liés à la scène sonore à étudier (nombre de sources, niveau de bruit, complexité) afin de choisir l'une ou l'autre des expressions désambiguïsées de la coordonnée sphérique d'azimut θ.

Selon un mode de réalisation, des première et deuxième coordonnées spatiales de la source sonore peuvent être des coordonnées sphériques d'azimut θ et d'élévation ϕ, les premier et deuxièmes capteurs peuvent faire partie d'un microphone ambisonique d'ordre 1 ou d'ordre supérieur à 1 plus généralement, les signaux issus du microphone ambisonique peuvent être un signal de pression $b_{00}^{1}(t)$ et trois signaux de gradient de pression $b_{11}^{1}(t)$, $b_{11}^{-1}(t)$ et $b_{10}^{1}(t)$.

Le premier signal, noté $s_{cardvirt1}$, et le second signal, noté $s_{cardvirt2}$, et un troisième signal $s_{cardvirt3}$ peuvent être obtenus à partir des signaux $b_{00}^{1}(t)$, $b_{11}^{1}(t)$, $b_{11}^{-1}(t)$ et $b_{10}^{1}(t)$ par :

$$s_{cardvirt1}(t,\theta,\phi) = \frac{b_{00}^{1}(t)}{2}\left(1+\frac{b_{11}^{-1}(t)}{b_{00}^{1}(t)}\right) ;$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{b_{00}^{1}(t)}{2}\left(1-\frac{b_{11}^{-1}(t)}{b_{00}^{1}(t)}\right) ;$$

$$s_{cardvirt3}(t,\theta,\phi) = \frac{b_{00}^{1}(t)}{2}\left(1+\frac{b_{10}^{1}(t)}{b_{00}^{1}(t)}\right) .$$

Les expressions des coordonnées sphériques d'azimut θ et d'élévation ϕ

peuvent alors être données par:

$$\phi = \sin^{-1}[2\frac{s_{cardvirt3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1]$$

$$\theta = \cos^{-1}[\frac{s^2_{cardvirt1}(t,\theta,\phi) - s^2_{cardvirt2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

avec $s_0 = b^1_{00}(t)$.

L'information supplémentaire peut être déterminée à partir d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux $b^1_{00}(t), b^1_{11}(t)$ et $b^{-1}_{11}(t)$.

Ainsi, la présente invention est applicable lorsque les signaux sonores sont initialement captés par un microphone ambisonique à l'ordre 1, tel que le microphone décrit dans le brevet US 4,042,779. En effet, en obtenant les premier et deuxième signaux, des microphones cardioïdes virtuels sont synthétisés, ce qui permet de revenir à des expressions semblables au premier mode de réalisation pour les coordonnées sphériques d'azimut et d'élévation. En revanche, dans le cas d'un microphone ambisonique à l'ordre 1, il n'est pas possible d'introduire un retard, et l'invention prévoit alors la prise en compte d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux captés par ces microphones, afin de lever l'incertitude sur la détermination de la coordonnée sphérique d'azimut. Ainsi, la direction de la source sonore peut être déterminée sans incertitude, sans toutefois nécessiter l'introduction de microphones supplémentaires.

En complément, l'information supplémentaire peut être une deuxième expression de la coordonnée sphérique d'azimut θ :

$$\theta(\omega) = \tan^{-1}(\frac{I_y(\omega)}{I_x(\omega)})$$

ω étant une pulsation du signal sonore émis par la source,

$I_y(\omega)$ étant donné par :

$$I_y(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[B^{1\,*}_{00}(\omega)B^{-1}_{11}(\omega)\right];$$

$I_x(\omega)$ étant donné par :

$$I_x(\omega) = \frac{1}{2\rho c} Re\left[B_{00}^{1\,*}(\omega) B_{11}^{1}(\omega)\right] ;$$

$B_{00}^{1}(\omega)$, $B_{11}^{1}(\omega)$ et $B_{11}^{-1}(\omega)$ désignant les transformées de Fourier des signaux $b_{00}^{1}(t)$, $b_{11}^{1}(t)$ et $b_{11}^{-1}(t)$ respectivement.

Ainsi, en obtenant une expression supplémentaire sur la coordonnée sphérique d'azimut θ, il est possible de lever l'incertitude liée à la première expression comprenant la fonction cosinus inverse. En effet, bien que la fonction tangente inverse présente également une incertitude, la fonction tangente inverse et la fonction cosinus inverse permettent d'obtenir deux estimations de la coordonnée sphérique d'azimut θ qui sont complémentaires. Par exemple, comme détaillé ultérieurement, l'utilisation d'un tableau permet de différencier quatre cas de figures, selon les intervalles dans lesquels se situent les deux estimations de la coordonnée sphérique d'azimut θ. Une valeur désambiguïsée de la coordonnée sphérique d'azimut θ peut être déterminée. En complément, il est possible de prendre en compte des facteurs liés à la scène sonore à étudier (nombre de sources, niveau de bruit, complexité) afin de choisir l'une ou l'autre des expressions désambiguïsées de la coordonnée sphérique d'azimut θ.

En complément ou en variante, une expression supplémentaire peut être déterminée pour la coordonnée sphérique d'élévation :

$$\phi(\omega) = \tan^{-1}\left(\frac{I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}}\right)$$

ω étant une pulsation du signal sonore émis par la source,

$I_y(\omega)$ étant la composante selon la coordonnée y du vecteur d'intensité acoustique du signal sonore et étant donnée par :

$$I_y(\omega) = \frac{1}{2\rho c} Re\left[B_{00}^{1\,*}(\omega) B_{11}^{-1}(\omega)\right] ;$$

$I_x(\omega)$ étant la composante selon la coordonnée x du vecteur d'intensité acoustique du signal sonore et étant donnée par :

$$I_x(\omega) = \frac{1}{2\rho c} Re\left[B_{00}^{1\,*}(\omega) B_{11}^{1}(\omega)\right] ;$$

278

$I_z(\omega)$ étant donné par :

$$I_z(\omega) = \frac{1}{2\rho c} Re\left[B_{00}^{1\,*}(\omega) B_{01}^{1}(\omega)\right] ;$$

$B_{00}^{1}(\omega)$, $B_{11}^{1}(\omega)$, $B_{01}^{1}(\omega)$ et $B_{11}^{-1}(\omega)$ désignant les transformées de Fourier des signaux $b_{00}^{1}(t)$, $b_{11}^{1}(t)$, $b_{01}^{1}(t)$ et $b_{11}^{-1}(t)$ respectivement.

La coordonnée sphérique d'élévation peut être déterminée à partir de l'expression ou de l'expression supplémentaire.

Ainsi, comme pour la détermination de la coordonnée sphérique d'azimut, il est possible de privilégier l'une ou l'autre des expressions déterminées pour la coordonnée sphérique d'élévation. A nouveau, ce choix peut être fait en fonction de facteurs liés à la scène sonore à étudier, tels que le nombre de sources, le niveau de bruit, la complexité, etc.

Un deuxième aspect de l'invention concerne un programme d'ordinateur comprenant des instructions de code de programme enregistrées sur un support lisible par un ordinateur, pour l'exécution des étapes du procédé selon le premier aspect de l'invention.

Un troisième aspect de l'invention concerne un dispositif de traitement de données pour la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le dispositif comprenant :

- une unité d'obtention d'au moins un premier signal et un deuxième signal à partir du signal sonore capté selon des directivités différentes par un premier capteur et un deuxième capteur ;
- une unité de déduction pour déduire des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, l'expression comportant une incertitude sur ladite coordonnée spatiale ;
- une première unité de détermination d'une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ;
- une seconde unité de détermination de la première coordonnée

spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

Un quatrième aspect de l'invention concerne un système d'acquisition de données sonores, comprenant une unité microphonique, l'unité microphonique comprenant au moins un premier capteur et un deuxième capteur aptes à capter des signaux sonores selon des directivités différentes, le système comprenant en outre un dispositif de traitement de données pour la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le dispositif comprenant:

- une unité d'obtention d'au moins un premier signal et un deuxième signal à partir du signal sonore capté par le premier capteur et le deuxième capteur ;
- une unité de déduction pour déduire des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, ladite expression comportant une incertitude sur la coordonnée spatiale ;
- une première unité de détermination d'une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ;
- une seconde unité de détermination de ladite première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

Selon un mode de réalisation, les premier et deuxième capteurs peuvent être des microphones cardioïdes. En variante, les premier et deuxième capteurs peuvent être des microphones bidirectionnels.

Un cinquième aspect de l'invention concerne un terminal de télécommunication comprenant un système d'acquisition de données sonores selon le quatrième mode de réalisation.

D'autres caractéristiques et avantages de l'invention apparaîtront à l'examen de la description détaillée ci-après, et des dessins annexés sur

lesquels:

- la figure 1 est un diagramme représentant les étapes générales d'un procédé de traitement de données selon un mode de réalisation;
- la figure 2 représente une structure générale d'un dispositif de traitement de données selon un mode de réalisation;
- la figure 3 illustre un système d'acquisition et de traitement de données sonores selon un mode de réalisation de l'invention ;
- la figure 4 illustre un terminal de télécommunication selon un mode de réalisation de l'invention ;
- la figure 5 illustre une unité microphonique selon un mode de réalisation de l'invention ;
- la figure 6 illustre les étapes du procédé selon un mode de réalisation de l'invention pour des signaux captés par l'unité microphonique de la figure 4 ;
- les figures 7a et 7b illustrent une unité microphonique selon un autre mode de réalisation de l'invention ;
- la figure 8 illustre les étapes du procédé selon un mode de réalisation de l'invention pour des signaux captés par l'unité microphonique des figures 6a et 6b.

La figure 1 est un diagramme illustrant les étapes générales d'un procédé de traitement de données selon un mode de réalisation de l'invention.

Le procédé permet la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel. On entend par coordonnée spatiale toute coordonnée parmi un système de trois coordonnées permettant de repérer la source sonore dans l'espace tridimensionnel. Aucune restriction n'est attachée au système de coordonnées considérées. Par exemple, il peut s'agir des coordonnées sphériques, cartésiennes ou cylindriques.

A une étape 10, au moins un premier signal et un deuxième signal sont obtenus à partir du signal sonore capté selon des directivités différentes par un premier capteur et un deuxième capteur. On entend par capteur tout système microphonique d'acquisition de données sonores. Les capteurs

considérés dépendent du système microphonique en question. De nombreux exemples de systèmes microphoniques sont présentés dans ce qui suit et l'invention s'applique ainsi à tout système microphonique. Les capteurs ayant des directivités différentes, ils captent deux signaux distincts, bien que ces signaux proviennent du même signal sonore émis par la source sonore.

A une étape 11, une expression d'au moins une première coordonnée spatiale de la source sonore est déterminée à partir des premier et deuxième signaux, une telle expression comportant une incertitude sur la coordonnée spatiale. Comme évoqué dans la partie introductive, l'incertitude peut être une ambiguïté angulaire de ±π/2 sur l'angle d'azimut. C'est par exemple le cas lorsque la première coordonnée spatiale est exprimée sous la forme d'une fonction cosinus inverse. La présente invention permet de lever une telle incertitude.

A cet effet, à une étape 12, une information supplémentaire relative à la première coordonnée spatiale de la source sonore est déterminée à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs. Comme détaillé dans ce qui suit, la comparaison peut être une différence entre les phases des signaux captés par les premier et deuxième capteurs ou un d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux captés.

A une étape 13, la première coordonnée spatiale de la source sonore est déterminée, avec certitude, sur la base de l'expression et de l'information supplémentaire. Ainsi, le procédé selon l'invention permet de lever l'incertitude sur la première coordonnée spatiale par l'utilisation d'une information supplémentaire déterminée à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs. La précision de la localisation de la source sonore est ainsi améliorée. Des exemples de procédés selon l'invention seront détaillés dans ce qui suit, en référence aux figures 5 et 7.

La figure 2 présente une structure générale d'un dispositif de traitement de données 20 selon un mode de réalisation de l'invention.

Le dispositif 20 comprend une unité d'obtention 21 d'au moins un premier signal et un deuxième signal à partir de signaux sonores captés selon des directivités différentes par un premier capteur et un deuxième capteur, ainsi qu'une unité de déduction 22 pour déduire des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, l'expression comportant une incertitude sur la coordonnée spatiale.

Le dispositif 20 comprend en outre une première unité de détermination 23 d'une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ainsi qu'une seconde unité de détermination 24 de la première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

La figure 3 illustre un système d'acquisition de signaux sonores selon un mode de réalisation de l'invention.

Le système comprend une unité microphonique 30 apte à capter des signaux sonores. Comme détaillé dans ce qui suit, l'unité microphonique 30 peut prendre diverses formes et comprendre plusieurs capteurs tels que des microphones cardioïdes et/ou bidirectionnels, ou qu'un microphone ambisonique. Aucune restriction n'est attachée à l'unité microphonique 30 considérée.

Le système d'acquisition comprend en outre le dispositif de traitement de données 20 décrit ci-avant. Dans ce qui suit, on fait l'hypothèse qu'il n'existe qu'une seule source sonore à chaque instant et par bande fréquentielle considérée. Ainsi, le traitement par le dispositif 20 s'effectue sur des fenêtres temporelles dont la taille est déterminée en fonction de l'écart des capteurs et en fonction d'un nombre d'échantillons fréquentiels souhaités. Selon l'invention, il est également possible d'ajouter des zéros (« *zeropadding* » en anglais) en fonction d'une discrétisation spectrale souhaitée.

Le système d'acquisition comprend en outre une unité d'encodage 32. A partir des directions des sources sonores déterminées grâce au dispositif de traitement de données 20, l'unité d'encodage 32 peut spatialiser des sources

virtuelles selon le type d'encodage spatial d'une technologie de restitution considérée. Par exemple, dans le cas d'un rendu binaural sur casque ou sur haut-parleurs, les directions des sources déterminent les HRTF à utiliser pour spatialiser les sons, avec une possibilité de choix personnalisé des HRTF pour l'auditeur. Toujours dans le cas d'un rendu binaural sur casque ou sur haut-parleurs, la correction de la position relative à la tête est possible grâce à l'utilisation d'un système de suivi des mouvements de la tête (« *head tracking* » en anglais). Dans un autre mode de restitution, l'unité d'encodage 32 synthétise des signaux ambisoniques aux différents ordres pour une diffusion sur casque ou sur un ensemble de haut-parleurs ad doc dont les positions sont connues.

La figure 4 illustre un terminal de télécommunication 40 selon un mode de réalisation de l'invention. Le terminal de télécommunication 40 peut être un téléphone mobile (de type Smartphone par exemple), un PDA (pour « *Personal Digital Assistant* ») ou encore une tablette tactile par exemple. Le terminal de télécommunication peut intégrer le système décrit sur la figure 3, et comprend à cet effet un ensemble de capteurs 41, correspondant à l'unité microphonique 30, un microprocesseur 42 et une mémoire 43.

La mémoire est apte à stocker un programme d'ordinateur comprenant des instructions de code de programme permettant l'exécution par le microprocesseur 42 des étapes du procédé selon l'invention. Le microprocesseur peut ainsi réaliser les fonctionnalités du dispositif 20, et éventuellement de l'unité d'encodage 32.

La figure 5 illustre une unité microphonique 30 selon un premier mode de réalisation de l'invention. L'unité microphonique 30 comprend trois capteurs qui sont des microphones cardioïdes 51, 52 et 53.

Les trois microphones 51, 52 et 53 sont présentés dans un plan x,z comprenant l'origine O, de l'espace orienté par un repère orthonormé comprenant les trois axes x, y et z.

Le premier microphone cardioïde 51 est dirigé selon l'axe x vers des valeurs croissantes tandis que le deuxième microphone cardioïde 52 est dirigé vers des valeurs de x décroissantes. Le troisième microphone 53 est dirigé

selon l'axe z vers des valeurs croissantes. Les directivités respectives des microphones 51, 52 et 53, en fonction des directions de pointage, sont illustrées par des cardioïdes 54, 55 et 56 vues dans le plan (x,z).

En effet, la fonction de directivité M d'un microphone cardioïde est exprimée par la relation :

$$M(\alpha) = \frac{1}{2}(1+\alpha) \qquad (1)$$

avec $\alpha = \vec{d}_s.\vec{d}_p$ (2)

où $\vec{d}_s$ est un vecteur définissant la direction de la source sonore et $\vec{d}_p$ le vecteur déterminant la direction de pointage du microphone.

Dans l'exemple de la figure 4, des directions de pointages $\vec{d}_{p1}, \vec{d}_{p2}$ et $\vec{d}_{p3}$ respectives des trois microphones 51, 52 et 53 peuvent être exprimées dans une base de coordonnées cartésiennes B_C :

$$\vec{d}_{p1} = \begin{vmatrix} 1 \\ 0 \\ 0 \end{vmatrix}_{B_C}, \quad \vec{d}_{p2} = \begin{vmatrix} -1 \\ 0 \\ 0 \end{vmatrix}_{B_C}, \quad \vec{d}_{p3} = \begin{vmatrix} 0 \\ 0 \\ 1 \end{vmatrix}_{B_C} \qquad (3)$$

Considérant que la direction de la source sonore est exprimée dans une base B_S en coordonnées sphériques ou dans la base de coordonnées cartésiennes B_C :

$$\vec{d}_s = \begin{vmatrix} \theta \\ \phi \\ r \end{vmatrix}_{B_S} = \begin{vmatrix} r.\cos\phi\cos\theta \\ r.\cos\phi\sin\theta \\ r.\sin\phi \end{vmatrix}_{B_C} \qquad (4)$$

où les coordonnées sphériques sont définies par le rayon r, l'angle d'azimut θ et l'angle d'élévation ϕ.

Des fonctions de directivité pour les trois microphones 51, 52 et 53 peuvent alors s'exprimer de la manière suivante :

$$M_{card1}(\theta,\phi) = \frac{1}{2}(1+r.\cos\phi\cos\theta) \qquad A$$

$$M_{card2}(\theta,\phi) = \frac{1}{2}(1-r.\cos\phi\cos\theta) \qquad B \qquad (5)$$

$$M_{card3}(\theta,\phi) = \frac{1}{2}(1 + r.\sin\phi) \qquad C$$

Par souci de simplification, il est considéré dans ce qui suit que r=1, ce qui ne modifie pas la direction de pointage.

La figure 6 illustre un premier mode de réalisation particulier d'un procédé selon l'invention, mis en œuvre lorsque l'unité microphonique 30 est l'unité représentée sur la figure 5.

La source sonore dont la direction est pointée par le vecteur $\vec{d}_s$ induit un signal $s_0(t)$ à l'origine O du repère. En considérant idéalement que les microphones 51, 52 et 53 sont placés à l'origine O, les signaux $s_{card1}(t)$, $s_{card2}(t)$ et $s_{card3}(t)$ captés respectivement par les microphones 51, 52 et 53 sont :

$$\begin{aligned} s_{card1}(t) &= M_{card1}(\theta,\phi)s_0(t) \qquad A \\ s_{card2}(t) &= M_{card2}(\theta,\phi)s_0(t) \qquad B \qquad (6) \\ s_{card3}(t) &= M_{card3}(\theta,\phi)s_0(t) \qquad C \end{aligned}$$

A une étape 60, des premier, deuxième et troisième signaux sont obtenus par l'unité d'obtention 21 à partir des signaux $s_{card1}(t)$, $s_{card2}(t)$ et $s_{card3}(t)$ captés respectivement par les microphones 51, 52 et 53. Dans cet exemple, les premier, deuxième et troisième signaux sont égaux aux signaux $s_{card1}(t)$, $s_{card2}(t)$ et $s_{card3}(t)$ respectivement.

A une étape 61, l'unité d'obtention peut appliquer un fenêtrage aux premier, deuxième et troisième signaux. De préférence, et afin de minimiser les oscillations dans le domaine fréquentiel, une trame temporelle est fenêtrée par une fenêtre à transition douce.

A une étape 62, l'unité d'obtention 21 applique une transformée de Fourier aux premier, deuxième et troisième signaux. Ainsi, les opérations décrites dans ce qui suit agissent dans le domaine fréquentiel, fréquence par fréquence.

Dans ce qui suit, certaines expressions sont encore données dans le domaine temporel : la transformée de Fourier étant linéaire, ces expressions seraient similaires dans le domaine fréquentiel (à une convention de notation près, et en remplaçant t par une pulsation).

Des relations de directivité 6A, 6B et 6C, l'unité de déduction 22 peut déduire à une étape 63 les expressions suivantes pour le signal acoustique $s_0(t)$ généré par la source sonore à l'origine O et pour les coordonnées spatiales θ et ϕ :

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi) \qquad (7)\ \text{(en combinant 5A, 5B et 6)}$$

$$\phi = \sin^{-1}[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1] \qquad (8)\ \text{(en combinant 5C et 6)}$$

$$\theta = \cos^{-1}[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}] \qquad (9)\ \text{(en combinant 5A et 5B)}$$

Ainsi des expressions des coordonnées spatiales θ et ϕ sont obtenues uniquement à partir de la directivité des microphones 51, 52 et 53 (les signaux captés respectivement par ces microphones).

L'expression de la coordonnée d'azimut θ prend toutefois la forme d'un cosinus inverse, et la coordonnée d'azimut θ est ainsi déterminée avec une incertitude de $\pm\pi/2$.

La présente invention permet de lever une telle incertitude et utilisant une information supplémentaire telle que précédemment décrite. L'information supplémentaire est déterminée par la première unité de détermination 23 à une étape 64.

Dans l'exemple illustré aux figures 5 et 6, l'information supplémentaire est le signe d'une différence entre les phases respectives (ou retard) des premier et second signaux.

En effet, un décalage entre des positions respectives des microphones 51 et 52, perpendiculaire à la direction de pointage de ces microphones et dans le plan (x,y) (donc selon l'axe y) est introduit selon l'invention.

Le signal s_{card1} capté par le premier microphone 51 décalé aux coordonnées cartésiennes (x_1,y_1,z_1) est décrit par :

$$s_{card1}(t) = M_{card1}(\theta,\phi)s_0(t-\tau_1) \qquad (10)$$

où τ_1 représente la différence de marche engendrée par la distance entre le microphone 51 et l'origine O, et qui peut s'exprimer par la relation :

$$\tau_1 = \frac{1}{c}\vec{d}_s.\vec{E}_1 \qquad (11)$$

où c est la célérité des ondes acoustiques et $\vec{E}_1$ est un vecteur déterminant l'emplacement du microphone 51 en coordonnées cartésiennes dans la base B_C :

$$\vec{E}_1 = \begin{vmatrix} x_1 \\ y_1 \\ z_1 \end{vmatrix}_{B_C} \qquad (12)$$

considérant que la direction $\vec{d}_s$ de la source sonore est exprimée dans la base sphérique B_S ou dans la base cartésienne B_C :

$$\vec{d}_s = \begin{vmatrix} \theta \\ \phi \\ r \end{vmatrix}_{B_S} = \begin{vmatrix} r.\cos\phi\cos\theta \\ r.\cos\phi\sin\theta \\ r.\sin\phi \end{vmatrix}_{B_C}$$

et qu'ainsi :

$$\tau_1 = \frac{1}{c}(x_1 r.\cos\phi\cos\theta + y_1 r.\cos\phi\sin\theta + z_1 r.\sin\phi) \qquad (13)$$

De même, on obtient pour le deuxième microphone 52 :

$$s_{card2}(t) = M_{card2}(\theta,\phi)s_0(t-\tau_2) \qquad (14)$$

avec $\tau_2 = \frac{1}{c}(x_2 r.\cos\phi\cos\theta + y_2 r.\cos\phi\sin\theta + z_2 r.\sin\phi)$ (15)

Dans le domaine fréquentiel, les signaux $s_{card1}(t)$ et $s_{card2}(t)$ deviennent $S_{card1}(\omega)$ et $S_{card2}(\omega)$, où $\omega = 2\pi f$ désigne la pulsation, f étant la fréquence associée au signal sonore émis par la source sonore.

Dans ce qui suit, la transformée de Fourier est notée FT[].

$$FT[s_{card1}(t,\theta,\phi)] = S_{card1}(\omega,\theta,\phi) = M_{card1}(\theta,\phi)S_0(\omega)e^{-j\omega\tau_1} \qquad (16)$$

$$FT[s_{card2}(t,\theta,\phi)] = S_{card2}(\omega,\theta,\phi) = M_{card2}(\theta,\phi)S_0(\omega)e^{-j\omega\tau_2} \qquad (17)$$

où $S_0(\omega) = |S_0(\omega)|e^{j\angle S_0(\omega)}$, '$\angle$'désignant la phase du signal sonore à l'origine O.

Ainsi :

$$S_{card1}(\omega,\theta,\phi) = M_{card1}(\theta,\phi)|S_0(\omega)|e^{j(\angle S_0(\omega)-\omega\tau_1)} \qquad (18)$$

$$S_{card2}(\omega,\theta,\phi) = M_{card2}(\theta,\phi)\left|S_0(\omega)\right|e^{j(\angle S_0(\omega)-\omega\tau_2)} \quad (19)$$

En notant $\angle S_1(\omega) = \angle S_0(\omega) - \omega\tau_1$ (20)

et $\angle S_2(\omega) = \angle S_0(\omega) - \omega\tau_2$ (21)

On obtient : $\angle S_1 - \angle S_2 = -\omega(\tau_1 - \tau_2)$ (22)

En notant $\tau_{12} = \tau_1 - \tau_2$ le retard temporel entre les signaux captés par les microphones 51 et 52, on obtient :

$$\tau_{12}(\omega) = -\frac{1}{\omega}(\angle S_1 - \angle S_2) \quad (23)$$

A une étape 65, la seconde unité de détermination 24 détermine la coordonnée spatiale θ sur la base de l'information supplémentaire (signe du retard ou différence de phase entre les signaux captés respectivement par les microphones 51 et 52) et de l'expression de θ comportant une incertitude (expression (9)).

Le retard temporel τ_{12} étant uniquement utilisé pour lever l'incertitude introduite par l'expression de la coordonnée d'azimut θ (expression (9)), seul le signe du retard temporel τ_{12} est utilisé en l'introduisant directement dans l'expression (9) :

$$\theta = \frac{\tau_{12}(\omega)}{\left|\tau_{12}(\omega)\right|}\cos^{-1}\left[\frac{S^2_{card1}(\omega,\theta,\phi) - S^2_{card2}(\omega,\theta,\phi)}{S_0(\omega)^2 \cdot \cos\phi}\right] \quad (24)$$

Les figures 7a et 7b illustrent une unité microphonique 30 selon un deuxième mode de réalisation de l'invention. L'unité microphonique 30 comprend trois capteurs, à savoir un premier microphone bidirectionnel 71, un second microphone bidirectionnel 73 et un microphone cardioïde 72. On entend par microphone bidirectionnel un microphone à gradient de pression.

Le premier microphone bidirectionnel 71 est placé sur l'axe x (voir figure 7a), le second microphone bidirectionnel 73 est placé sur l'axe y (voir figure 7b) et le microphone cardioïde 72 est orienté sur l'axe z vers les valeurs croissantes (voir figures 7a et 7b).

La directivité, en fonction des directions de pointage, du premier

microphone bidirectionnel 71 est représentée sous forme de deux sphères orientées vers les x positifs et x négatifs et présentées dans le plan (x,z) sous les références 74.1 et 74.2 de la figure 7a, respectivement.

La directivité, en fonction des directions de pointage, du second microphone bidirectionnel 73 est représentée sous forme de deux sphères orientées vers les y positifs et y négatifs et présentées dans le plan (y,z) sous les références 76.1 et 76.2 de la figure 7b, respectivement.

La directivité du microphone cardioïde 72, en fonction des directions de pointage, est illustrée par une cardioïde 75 vue dans le plan (x,z) sur la figure 7a et dans le plan (y,z) sur la figure 7b.

La figure 8 illustre un deuxième mode de réalisation particulier d'un procédé selon l'invention, mis en œuvre lorsque l'unité microphonique 30 est l'unité représentée sur les figures 7a et 7b.

L'unité microphonique 30 permet ainsi l'obtention de signaux issus d'une paire de microphones bidirectionnels 71 et 73 placés perpendiculairement dans le plan horizontal (x,y). L'invention propose alors de reconstruire de façon virtuelle les signaux captés par les microphones cardioïdes 51 et 52 de la figure 5 afin d'obtenir des premier et seconds signaux comparables à ceux obtenus à l'étape 50 de la figure 5.

A cet effet, à une étape 80, l'unité d'obtention 21 détermine des premier et seconds signaux à partir du signal sonore s_{bi1} capté par le premier microphone bidirectionnel 71 et du signal sonore s_{bi2} capté par le second microphone bidirectionnel 73.

Les expressions des signaux s_{bi1} et s_{bi2}, et du signal s_{card3} capté par le microphone cardioïde 72, sont données par les formules suivantes :

$$s_{bi1}(t,\theta,\phi) = s_0(t)\cos\theta\cos\phi \quad \text{A}$$

$$s_{bi2}(t,\theta,\phi) = s_0(t)\sin\theta\cos\phi \quad \text{B} \qquad (25)$$

$$s_{card3}(t,\theta,\phi) = \frac{s_0(t)}{2}(1+\sin\phi) \quad \text{C}$$

290

Le premier signal $s_{cardvirt1}$ et le second signal $s_{cardvirt2}$ qui auraient été captés par deux microphones cardioïdes sont reconstruits de la manière suivante :

$$s_{cardvirt1}(t,\theta,\phi) = \frac{s_0(t)}{2}(1 + \frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \qquad (26)$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{s_0(t)}{2}(1 - \frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \qquad (27)$$

$$\text{avec } s_0(t) = \frac{s_{bi1}^2(t,\theta,\phi) + s_{bi2}^2(t,\theta,\phi) + 4s_{card3}^2(t,\theta,\phi)}{4s_{card3}(t,\theta,\phi)} \qquad (28)$$

L'unité d'obtention 21 obtient ainsi des premier, deuxième et troisième signaux $s_{cardvirt1}$, $s_{cardvirt2}$ et s_{card3}.

A une étape 81, l'unité d'obtention 21 peut appliquer un fenêtrage aux premier, deuxième et troisième signaux. Comme précédemment expliqué, afin de minimiser les oscillations dans le domaine fréquentiel, une trame temporelle est fenêtrée par une fenêtre à transition douce.

A une étape 82, l'unité d'obtention 21 applique une transformée de Fourier aux premier, deuxième et troisième signaux. Ainsi, les opérations décrites dans ce qui suit agissent dans le domaine fréquentiel, fréquence par fréquence. A nouveau, certaines expressions sont encore données dans le domaine temporel : la transformée de Fourier étant linéaires, elles seraient similaires dans le domaine fréquentiel.

A une étape 83, l'unité de déduction 22 peut déduire des premier, deuxième et troisième signaux les expressions suivantes pour les coordonnées spatiales θ et ϕ :

$$\phi = \sin^{-1}[2\frac{s_{card3}(t,\theta,\phi)}{s_O(t,\theta,\phi)} - 1] \qquad (29)$$

$$\theta = \cos^{-1}[\frac{s_{cardvirt1}^2(t,\theta,\phi) - s_{cardvirt2}^2(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}] \qquad (30)$$

A nouveau, l'expression de la coordonnée spatiale θ présente une incertitude due à la fonction cosinus inverse. Dans une telle configuration virtuelle, l'incertitude précitée ne peut être levée en introduisant un retard entre les microphones bidirectionnels 71 et 73.

Toutefois, une information supplémentaire relative à d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux captés par les premier et second microphones bidirectionnels 71 et 73, peut être déterminée par la première unité de détermination 23, afin de lever l'incertitude sur l'expression (30).

L'intensité acoustique active est un vecteur lié à la pression et à la vitesse acoustique particulaire par la relation suivante, donnée dans le domaine fréquentiel :

$$I(\omega) = \begin{vmatrix} I_x(\omega) \\ I_y(\omega) \\ I_z(\omega) \end{vmatrix} = \frac{1}{2}\mathrm{Re}\left\{ P^*(\omega) \begin{vmatrix} V_x(\omega) \\ V_y(\omega) \\ V_z(\omega) \end{vmatrix} \right\} \qquad (31)$$

où $P^*(\omega)$ correspond au conjugué de la pression acoustique et les trois signaux $V_x(\omega)$, $V_y(\omega)$ et $V_z(\omega)$ représentent les trois composantes du vecteur de vitesse particulaire.

On considère le cas d'une onde plane progressive dont la pression est décrite par la relation suivante (dans laquelle $\vec{k}$ définit le vecteur d'onde) :

$$P(\omega, r, \theta, \phi) = P_0(\omega) e^{-j\vec{k}.\vec{r}} \qquad (32)$$

La vitesse particulaire se déduit de la relation d'Euler :

$$\vec{V}(\omega, r, \theta, \phi) = \frac{P(\omega, r, \theta, \phi)}{\rho c} \frac{\vec{k}}{\left|\vec{k}\right|} \qquad (33)$$

où ρ est la masse volumique du milieu de propagation et c la célérité des ondes acoustiques.

Ainsi, l'intensité acoustique est donnée par :

$$I(\omega) = \begin{vmatrix} I_x(\omega) = \dfrac{\left|P_0(\omega)\right|^2}{2\rho c} \dfrac{k_x}{\left|\vec{k}\right|} \\ I_y(\omega) = \dfrac{\left|P_0(\omega)\right|^2}{2\rho c} \dfrac{k_y}{\left|\vec{k}\right|} \\ I_z(\omega) = \dfrac{\left|P_0(\omega)\right|^2}{2\rho c} \dfrac{k_z}{\left|\vec{k}\right|} \end{vmatrix} \qquad (34)$$

292

Le vecteur intensité $\mathrm{I}(\omega)$ est colinéaire au vecteur d'onde, c'est-à-dire que sa direction est identique à la direction de propagation du signal sonore. La direction du vecteur d'intensité $\mathrm{I}(\omega)$ permet donc d'accéder à une estimation de la direction de la source sonore.

La projection V_{xy} sur le plan horizontal (x,y) de la vitesse particulaire est exprimée par :

$$V_{xy}(\omega) = \frac{1}{\rho c}\left[X(\omega)\vec{e}_x + Y(\omega)\vec{e}_y\right] \qquad (35)$$

où X et Y sont les signaux captés par les microphones bidirectionnels 71 et 73, respectivement.

Les signaux associés à la pression et aux composantes de la vitesse particulaire sont obtenus par la relation :

$$P_0(\omega) = S_0(\omega)$$
$$X(\omega) = S_{bi1}(\omega)$$
$$Y(\omega) = S_{bi2}(\omega)$$

Les composantes de l'intensité acoustique dans le plan (x,y) s'en déduisent de la façon suivante :

$$I_x(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[S_0^*(\omega)S_{bi1}(\omega)\right] \qquad (36)$$

$$I_y(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[S_0^*(\omega)S_{bi2}(\omega)\right] \qquad (37)$$

Le vecteur intensité étant colinéaire au vecteur d'onde, la tangente inverse du rapport entre les composantes de l'intensité acoustique des expressions (36) et (37) donne une estimation de la coordonnée spatiale θ et ainsi :

$$\theta(\omega) = \tan^{-1}\left(\frac{I_y(\omega)}{I_x(\omega)}\right) \qquad (38)$$

L'information supplémentaire $\frac{I_y(\omega)}{I_x(\omega)}$ est reliée à la coordonnée spatiale θ par une fonction tangente inverse ce qui introduit une incertitude droite-gauche, qui est complémentaire de l'incertitude due au cosinus inverse dans l'expression (30).

La seconde unité de détermination peut alors, à une étape 85, utiliser de façon conjointe l'information supplémentaire et l'expression (30) afin de déterminer avec certitude la coordonnée spatiale θ.

A cet effet, le tableau 1 ci-dessous illustre comment lever l'incertitude sur la coordonnée spatiale θ.

	θ **Réel**	θ **estimé**		**Opération à réaliser**	
		directivité	**intensité**	**directivité**	**intensité**
Cas	$\left[-\pi,-\frac{\pi}{2}\right]$	$\left[\frac{\pi}{2},\pi\right]$	$\left[0,\frac{\pi}{2}\right]$	$-\theta$	$\theta-\pi$
	$\left[-\frac{\pi}{2},0\right]$	$\left[0,\frac{\pi}{2}\right]$	$\left[-\frac{\pi}{2},0\right]$	$-\theta$	θ
	$\left[0,\frac{\pi}{2}\right]$	$\left[0,\frac{\pi}{2}\right]$	$\left[0,\frac{\pi}{2}\right]$	θ	θ
	$\left[\frac{\pi}{2},\pi\right]$	$\left[\frac{\pi}{2},\pi\right]$	$\left[-\frac{\pi}{2},0\right]$	θ	$\theta+\pi$

Tableau 1

Les colonnes 2 et 3 du tableau 1 (regroupées sous « θ estimé ») décrivent les différents cas de figure à l'issue des étapes 83 et 84, respectivement. La valeur réelle de la coordonnée spatiale θ est donnée par la première colonne (« θ réel »).

Les opérations à appliquer aux valeurs de la coordonnée spatiale θ estimées à l'issue des étapes 83 et 84 sont décrites dans les colonnes 4 et 5 (regroupées sous « Opération à réaliser »). Théoriquement, les coordonnées spatiales θ obtenues en effectuant les opérations des colonnes 4 et 5 sont les mêmes. Cependant, en pratique, en raison de la scène sonore à étudier et des conditions d'enregistrement (nombre de sources, niveau de bruit, complexité, etc), l'application de l'une ou l'autre des opérations décrites aux colonnes 4 et 5

294

peut donner une meilleure estimation de la coordonnée spatiale θ et est donc à privilégier.

Selon un troisième mode de réalisation spécifique de l'invention, l'unité microphonique 30 peut être un microphone ambisonique d'ordre 1, tel que le microphone décrit dans le brevet US 4,042,779 introduit précédemment. Plus généralement, l'unité microphonique peut être tout microphone ambisonique d'ordre supérieur à 1.

Un microphone ambisonique d'ordre 1 ou d'ordre supérieur à 1 plus généralement est apte à délivrer quatre signaux $b_{00}^{1}(t), b_{11}^{1}(t), b_{11}^{-1}(t)$ et $b_{10}^{1}(t)$ (dans le domaine temporel). Le signal $b_{00}^{1}(t)$ représente le signal de pression, tandis que les signaux $b_{11}^{1}(t), b_{11}^{-1}(t)$ et $b_{10}^{1}(t)$ correspondent à trois microphones bidirectionnels selon les axes x, y et z respectivement.

Comme dans le second mode de réalisation présenté précédemment, les signaux captés par l'unité microphonique 30 sont utilisés par l'unité d'obtention 30 afin de synthétiser un dispositif microphonique virtuel en déduisant des premier, deuxième et troisième signaux correspondant à trois microphones cardioïdes virtuels.

Les premier, deuxième et troisième signaux, notés respectivement $s_{cardvirt1}$, $s_{cardvirt2}$ et $s_{cardvirt3}$ sont obtenus par l'unité d'obtention 21 à partir des signaux ambisoniques à l'ordre 1 de la façon suivante :

$$s_{cardvirt1}(t,\theta,\phi) = \frac{b_{00}^{1}(t)}{2}\left(1+\frac{b_{11}^{-1}(t)}{b_{00}^{1}(t)}\right) \qquad (39)$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{b_{00}^{1}(t)}{2}\left(1-\frac{b_{11}^{-1}(t)}{b_{00}^{1}(t)}\right) \qquad (40)$$

$$s_{cardvirt3}(t,\theta,\phi) = \frac{b_{00}^{1}(t)}{2}\left(1+\frac{b_{10}^{1}(t)}{b_{00}^{1}(t)}\right) \qquad (41)$$

Tout comme expliqué précédemment, l'unité d'obtention 21 peut fenêtrer les premier, deuxième et troisième signaux et leur appliquer une transformée de Fourier afin de passer dans le domaine fréquentiel.

Les expressions suivantes pour le signal acoustique $s_0(t)$ généré par la source sonore à l'origine O et pour les coordonnées spatiales θ et ϕ sont alors obtenues par l'unité de déduction 22:

$$s_0(t) = b_{00}^1(t) \qquad (42)$$

$$\phi = \sin^{-1}[2\frac{s_{virtcard3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1] \qquad (43)$$

$$\theta = \cos^{-1}[\frac{s_{virtcard1}^2(t,\theta,\phi) - s_{virtcard2}^2(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}] \qquad (44)$$

A nouveau, l'expression (44) de la coordonnée spatiale θ présente une incertitude. Comme dans le deuxième mode de réalisation présenté précédemment, cette incertitude peut être levée en exploitant l'information supplémentaire liée à l'intensité acoustique dérivée des signaux du format B.

A cet effet, la première unité de détermination 23 détermine l'information supplémentaire liée à l'intensité acoustique dérivée des signaux du format B.

Les trois composantes de l'intensité acoustique sont calculées de la façon suivante :

$$\mathrm{I_x}(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[\mathrm{B_{00}^1}^*(\omega)\mathrm{B_{11}^1}(\omega)\right] \qquad (45)$$

$$\mathrm{I_y}(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[\mathrm{B_{00}^1}^*(\omega)\mathrm{B_{11}^{-1}}(\omega)\right] \qquad (46)$$

$$\mathrm{I_z}(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[\mathrm{B_{00}^1}^*(\omega)\mathrm{B_{01}^1}(\omega)\right] \qquad (47)$$

Il est alors possible de déterminer la direction de la source sonore (coordonnées spatiales θ et ϕ) grâce aux relations suivantes :

$$\theta(\omega) = \tan^{-1}\left(\frac{\mathrm{I_y}(\omega)}{\mathrm{I_x}(\omega)}\right) \qquad (48)$$

$$\phi(\omega) = \tan^{-1}\left(\frac{\mathrm{I_z}(\omega)}{\sqrt{\mathrm{I_x^2}(\omega) + \mathrm{I_y^2}(\omega)}}\right) \qquad (49)$$

Ainsi, comme dans le deuxième mode de réalisation, la seconde unité de détermination 24 peut utiliser l'expression (44) et l'information

supplémentaire de l'expression (48) pour déterminer la coordonnée spatiale θ. A cet effet, le Tableau 1 détaillé précédemment peut être utilisé.

On remarque en outre que dans ce troisième mode de réalisation, la coordonnée spatiale d'élévation ϕ est déterminée à la fois par l'expression (43) et par l'expression (49). Ainsi, comme pour la coordonnée spatiale θ, il est possible de privilégier l'une ou l'autre des méthodes d'estimation en fonction de la scène sonore à étudier (nombre de sources, niveau de bruit, complexité, etc) afin d'obtenir une meilleure localisation de la source sonore.

Selon un quatrième mode de réalisation spécifique de l'invention, l'unité microphonique 30 peut comprendre uniquement deux capteurs, qui sont des microphones cardioïdes (par exemple uniquement les microphones 51 et 52 de la figure 5).

Ainsi, un premier microphone cardioïde est dirigé selon l'axe x vers des valeurs positives et le second microphone cardioïde est dirigé selon l'axe x vers des valeurs négatives.

Les directivités des premier et deuxième microphones cardioïdes sont données par les expressions (5A) et (5B), et les signaux captés par ces deux microphones (qui, comme dans le premier mode de réalisation, sont également les premier et second signaux obtenus par l'unité d'obtention 21) sont donnés par les expressions (6A) et (6B). Afin de déterminer la coordonnée spatiale d'azimut θ, l'expression (9) est utilisée en fixant la coordonnée spatiale d'élévation ϕ à une valeur arbitraire ϕ_0. De préférence, la coordonnée spatiale d'élévation ϕ est fixée au plus proche du plan horizontal (valeur de ϕ faible), afin de minimiser l'erreur de localisation.

L'incertitude sur la coordonnée spatiale d'azimut θ due à l'expression (9) est résolue en décalant les premier et deuxième microphones cardioïdes sur l'axe y de façon à introduire un retard de la même manière que dans la premier mode de réalisation. L'information supplémentaire de l'expression (23) est alors utilisée afin d'obtenir l'expression (24) de la coordonnée spatiale d'azimut θ, en fixant ϕ à ϕ_0.

Selon un cinquième mode de réalisation spécifique de l'invention, l'unité microphonique 30 peut comprendre uniquement deux capteurs, qui sont des microphones bidirectionnels (par exemple uniquement les microphones 71 et 73 des figures 7a et 7b).

Dans ce cinquième mode de réalisation, un dispositif microphonique virtuel est synthétisé par l'unité d'obtention 21 afin d'obtenir des premier et second signaux à partir des signaux captés par les microphones bidirectionnels (expressions (25A) et (25B)), les premier et second signaux étant sous la forme de signaux captés par des microphones cardioïdes (expressions (26) et (27)).

L'expression (28), quant à elle, est approximée par l'expression suivante :

$$s_0(t) = \sqrt{s_{bi1}^2(t,\theta,\phi) + s_{bi2}^2(t,\theta,\phi)} \qquad (50)$$

A titre illustratif, cinq modes de réalisation ont été présentés ci-avant. Bien entendu, la présente invention ne se limite pas à ces exemples et s'étend à d'autres variantes, en fonction notamment des capteurs de l'unité microphonique 30.

Les résultats issus des cinq modes de réalisation présentés ci-avant permettent d'obtenir des estimations à chaque instant et pour chaque bande de fréquence d'au moins la coordonnée spatiale θ. Afin d'éviter les sauts intempestifs des sources sonores dus aux erreurs de localisation, il est possible d'effectuer un lissage des résultats obtenus dans le domaine fréquentiel et dans le domaine temporel.

REVENDICATIONS

1. Procédé de traitement de données pour la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le procédé comprenant les étapes suivantes :

- obtenir (10) au moins un premier signal et un deuxième signal à partir du signal sonore capté selon des directivités différentes par un premier capteur (51;71) et un deuxième capteur (52 ;73) ;
- déduire (11) des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, ladite expression comportant une incertitude sur ladite coordonnée spatiale ;
- déterminer (12) une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ;
- déterminer (13) ladite première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

2. Procédé selon la revendication 1, dans lequel, l'espace étant orienté selon trois axes x, y et z, les premier et deuxième capteurs sont des microphones cardioïdes (51 ;52), le premier microphone cardioïde étant situé à une première position de l'espace et orienté selon l'axe x dans un sens croissant et le deuxième microphone cardioïde étant situé à une deuxième position de l'espace et orienté selon l'axe x dans un sens décroissant, le procédé comprenant initialement :

- modifier la première ou la deuxième position en vue d'introduire un décalage selon l'axe y entre le premier microphone cardioïde et le deuxième microphone cardioïde,

dans lequel les premier et deuxième signaux correspondent aux signaux captés respectivement par les premier et deuxième microphones cardioïdes, et dans lequel l'information supplémentaire est le signe d'une différence entre des phases respectives des premier et seconds signaux.

3. Procédé selon la revendication 2, dans lequel un troisième signal s_{card3} capté par un troisième microphone cardioïde (53) orienté selon l'axe z dans un sens croissant est obtenu, dans lequel des première et deuxième coordonnées spatiales de la source sonore sont respectivement les coordonnées sphériques d'azimut θ et d'élévation ϕ et dans lequel, le premier signal capté par le premier microphone cardioïde (51) étant noté s_{card1} et le second signal capté par le second microphone cardioïde (52) étant noté s_{card2}, les expressions des coordonnées sphériques d'azimut θ et d'élévation ϕ sont données par :

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi)$$

$$\phi = \sin^{-1}[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1]$$

$$\theta = \cos^{-1}[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

et dans lequel le signe de la coordonnée sphérique d'azimut θ est donné par la différence de phases entre les premier et deuxième signaux.

4. Procédé selon la revendication 2, dans lequel des première et deuxième coordonnées spatiales de la source sonore sont respectivement les coordonnées sphériques d'azimut θ et d'élévation ϕ, dans lequel, le premier signal capté par le premier microphone cardioïde (51) étant noté s_{card1} et le second signal capté par le second microphone cardioïde (52) étant noté s_{card2}, l'expression de la coordonnée sphérique d'azimut θ est donnée par :

$$s_0(t) = s_{card1}(t,\theta,\phi) + s_{card2}(t,\theta,\phi)$$

$$\theta = \cos^{-1}[\frac{s^2_{card1}(t,\theta,\phi) - s^2_{card2}(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

dans lequel la coordonnée sphérique d'élévation ϕ est fixée arbitrairement et dans lequel le signe de la coordonnée sphérique d'azimut θ est donné par la différence de phases entre les premier et deuxième signaux.

5. Procédé selon la revendication 1, dans lequel, l'espace étant orienté selon trois axes x, y et z, des première et deuxième coordonnées spatiales de la source sonore sont les coordonnées sphériques d'azimut θ et d'élévation ϕ, les premier et deuxième capteurs sont des capteurs bidirectionnels (71 ;73), ledit premier capteur étant orienté selon l'axe x et captant le signal noté s_{bi1} et ledit deuxième capteur étant orienté selon l'axe y et captant le signal noté s_{bi2},

dans lequel un troisième capteur cardioïde (72) est dirigé selon l'axe z croissant et apte à capter un signal noté s_{card3},

dans lequel les premiers et second signaux sont notés respectivement $s_{cardvirt1}$ et $s_{cardvirt2}$ et sont obtenus par:

$$s_{cardvirt1}(t,\theta,\phi) = \frac{s_0(t)}{2}(1+\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \ ;$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{s_0(t)}{2}(1-\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \ ;$$

$$\text{avec } s_0(t) = \frac{s_{bi1}^2(t,\theta,\phi)+s_{bi2}^2(t,\theta,\phi)+4s_{card3}^2(t,\theta,\phi)}{4s_{card3}^2(t,\theta,\phi)} \ ;$$

dans lequel les expressions des coordonnées sphériques d'azimut θ et d'élévation ϕ sont données par:

$$\phi = \sin^{-1}[2\frac{s_{card3}(t,\theta,\phi)}{s_0(t,\theta,\phi)}-1]$$

$$\theta = \cos^{-1}[\frac{s_{cardvirt1}^2(t,\theta,\phi)-s_{cardvirt2}^2(t,\theta,\phi)}{s_0(t)^2\cdot\cos\phi}]$$

dans lequel l'information supplémentaire est déterminée à partir d'une direction d'un vecteur d'intensité acoustique du signal sonore, ladite direction étant calculée à partir des signaux s_{bi1} et s_{bi2}.

6. Procédé selon la revendication 1, dans lequel, l'espace étant orienté selon trois axes x, y et z, les première et deuxième coordonnées spatiales sont les coordonnées sphériques d'azimut θ et d'élévation ϕ, les premier et

deuxième capteurs sont des capteurs bidirectionnels (71 ;73), ledit premier capteur étant orienté selon l'axe x et captant le signal noté s_{bi1} et ledit deuxième capteur étant orienté selon l'axe y et captant le signal noté s_{bi2},

dans lequel les premiers et second signaux sont notés respectivement $s_{cardvirt1}$ et $s_{cardvirt2}$ et sont obtenus de la manière suivante :

$$s_{cardvirt1}(t,\theta,\phi) = \frac{s_0(t)}{2}(1+\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \ ;$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{s_0(t)}{2}(1-\frac{s_{bi2}(t,\theta,\phi)}{s_0(t)}) \ ;$$

avec $s_0(t) = \sqrt{s_{bi1}^2(t,\theta,\phi) + s_{bi2}^2(t,\theta,\phi)}$;

dans lequel l'expression de la coordonnée sphérique d'azimut θ est donnée par:

$$\theta = \cos^{-1}[\frac{s_{cardvirt1}^2(t,\theta,\phi) - s_{cardvirt2}^2(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

dans lequel la coordonnée sphérique d'élévation ϕ est fixée arbitrairement ;

dans lequel l'information supplémentaire est déterminée à partir d'une direction d'un vecteur d'intensité acoustique du signal sonore, ladite direction étant calculée à partir des signaux s_{bi1} et s_{bi2}.

7. Procédé selon l'une des revendications 5 et 6, dans lequel l'information supplémentaire est une deuxième expression de la coordonnée d'azimut θ :

$$\theta(\omega) = \tan^{-1}(\frac{I_y(\omega)}{I_x(\omega)})$$

ω étant une pulsation du signal sonore émis par la source,

$I_y(\omega)$ étant la composante selon la coordonnée y du vecteur d'intensité acoustique du signal sonore, donnée par :

$$I_y(\omega) = \frac{1}{2\rho c}\mathrm{Re}[S_0^*(\omega)S_{bi1}(\omega)] \ ;$$

$I_x(\omega)$ étant la composante selon la coordonnée x du vecteur d'intensité acoustique du signal sonore donnée par :

302 $$I_x(\omega) = \frac{1}{2\rho c}\mathrm{Re}[S_0^*(\omega)S_{bi2}(\omega)] \ ;$$

$S_0(\omega)$, $S_{bi1}(\omega)$ et $S_{bi2}(\omega)$ désignant les transformées de Fourier des signaux $s_0(t)$, $s_{bi}1(t)$ et $s_{bi2}(t)$ respectivement.

8. Procédé selon la revendication 1, dans lequel des première et deuxième coordonnées spatiales de la source sonore sont des coordonnées sphériques d'azimut θ et d'élévation ϕ, les premier et deuxièmes capteurs font partie d'un microphone ambisonique, dans lequel, les signaux issus du microphone ambisonique sont un signal de pression $b_{00}^1(t)$ et trois signaux de gradient de pression $b_{11}^1(t)$, $b_{11}^{-1}(t)$ et $b_{10}^1(t)$;

dans lequel le premier signal, noté $s_{cardvirt1}$, et le second signal, noté $s_{cardvirt2}$, et un troisième signal $s_{cardvirt3}$ sont obtenus à partir des signaux $b_{00}^1(t), b_{11}^1(t), b_{11}^{-1}(t)$ et $b_{10}^1(t)$ par :

$$s_{cardvirt1}(t,\theta,\phi) = \frac{b_{00}^1(t)}{2}(1+\frac{b_{11}^{-1}(t)}{b_{00}^1(t)}) \ ;$$

$$s_{cardvirt2}(t,\theta,\phi) = \frac{b_{00}^1(t)}{2}(1-\frac{b_{11}^{-1}(t)}{b_{00}^1(t)}) \ ;$$

$$s_{cardvirt3}(t,\theta,\phi) = \frac{b_{00}^1(t)}{2}(1+\frac{b_{10}^1(t)}{b_{00}^1(t)}) \ ;$$

dans lequel les expressions des coordonnées sphériques d'azimut θ et d'élévation ϕ sont données par:

$$\phi = \sin^{-1}[2\frac{s_{cardvirt3}(t,\theta,\phi)}{s_0(t,\theta,\phi)} - 1]$$

$$\theta = \cos^{-1}[\frac{s_{cardvirt1}^2(t,\theta,\phi) - s_{cardvirt2}^2(t,\theta,\phi)}{s_0(t)^2 \cdot \cos\phi}]$$

avec $s_0 = b_{00}^1(t)$;

dans lequel l'information supplémentaire est déterminée à partir d'une direction d'un vecteur d'intensité acoustique du signal sonore, la direction étant calculée à partir des signaux $b_{00}^1(t), b_{11}^1(t)$ et $b_{11}^{-1}(t)$.

9. Procédé selon la revendication 8, dans lequel l'information supplémentaire est une deuxième expression de la coordonnée sphérique

d'azimut θ :

$$\theta(\omega) = \tan^{-1}\left(\frac{I_y(\omega)}{I_x(\omega)}\right)$$

ω étant une pulsation du signal sonore émis par la source,

$I_y(\omega)$ étant la composante selon la coordonnée y du vecteur d'intensité acoustique du signal sonore et étant donnée par :

$$I_y(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[B_{00}^{1\,*}(\omega)B_{11}^{-1}(\omega)\right]$$

$I_x(\omega)$ étant la composante selon la coordonnée x du vecteur d'intensité acoustique du signal sonore et étant donnée par :

$$I_x(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[B_{00}^{1\,*}(\omega)B_{11}^{1}(\omega)\right];$$

$B_{00}^{1}(\omega)$, $B_{11}^{1}(\omega)$ et $B_{11}^{-1}(\omega)$ désignant les transformées de Fourier des signaux $b_{00}^{1}(t)$, $b_{11}^{1}(t)$ et $b_{11}^{-1}(t)$ respectivement.

10. Procédé selon la revendication 8 ou 9, dans lequel une expression supplémentaire est déterminée pour la coordonnée sphérique d'élévation :

$$\phi(\omega) = \tan^{-1}\left(\frac{I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}}\right)$$

ω étant une pulsation du signal sonore émis par la source,

$I_y(\omega)$ étant donné par :

$$I_y(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[B_{00}^{1\,*}(\omega)B_{11}^{-1}(\omega)\right];$$

$I_x(\omega)$ étant donné par :

$$I_x(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[B_{00}^{1\,*}(\omega)B_{11}^{1}(\omega)\right];$$

$I_z(\omega)$ étant donné par :

$$I_z(\omega) = \frac{1}{2\rho c}\mathrm{Re}\left[B_{00}^{1\,*}(\omega)B_{01}^{1}(\omega)\right]$$

304 $B_{00}^{1}(\omega)$, $B_{11}^{1}(\omega)$, $B_{01}^{1}(\omega)$ et $B_{11}^{-1}(\omega)$ désignant les transformées de Fourier des

signaux $b_{00}^{1}(t)$, $b_{11}^{1}(t)$, $b_{01}^{1}(t)$ et $b_{11}^{-1}(t)$ respectivement ;

et dans lequel la coordonnée sphérique d'élévation est déterminée à partir de ladite expression ou de ladite expression supplémentaire.

11. Produit programme d'ordinateur comprenant des instructions de code de programme enregistrées sur un support lisible par un ordinateur, pour l'exécution des étapes du procédé selon l'une quelconque des revendications 1 à 10.

12. Dispositif de traitement de données pour la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le dispositif comprenant :

- une unité d'obtention (21) d'au moins un premier signal et un deuxième signal à partir du signal sonore capté selon des directivités différentes par un premier capteur et un deuxième capteur ;
- une unité de déduction (22) pour déduire des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, ladite expression comportant une incertitude sur ladite coordonnée spatiale ;
- une première unité de détermination (23) d'une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ;
- une seconde unité de détermination (24) de ladite première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

13. Système d'acquisition de données sonores, comprenant une unité microphonique (30), ladite unité microphonique comprenant au moins un premier capteur (51 ;71) et un deuxième capteur (52 ;73) aptes à capter des signaux sonores selon des directivités différentes, ledit système comprenant en outre un dispositif de traitement de données (20) pour la

détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le dispositif comprenant:

- une unité d'obtention (21) d'au moins un premier signal et un deuxième signal à partir du signal sonore capté par le premier capteur et le deuxième capteur ;
- une unité de déduction (22) pour déduire des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, ladite expression comportant une incertitude sur ladite coordonnée spatiale ;
- une première unité de détermination (23) d'une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ;
- une seconde unité de détermination (24) de ladite première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

14. Système selon la revendication 13, dans lequel les premier et deuxième capteurs sont des microphones cardioïdes (51 ;52).

15. Terminal de télécommunication comprenant un système selon la revendication 13.

306

ABREGE

Acquisition de données sonores spatialisées

L'invention concerne un procédé de traitement de données pour la détermination d'au moins une coordonnée spatiale d'une source sonore émettant un signal sonore, dans un espace tridimensionnel, le procédé comprenant les étapes suivantes :

- obtenir (10) au moins un premier signal et un deuxième signal à partir du signal sonore capté selon des directivités différentes par un premier capteur et un deuxième capteur ;
- déduire (11) des premier et deuxième signaux une expression d'au moins une première coordonnée spatiale de la source sonore, l'expression comportant une incertitude sur ladite coordonnée spatiale ;
- déterminer (12) une information supplémentaire relative à la première coordonnée spatiale de la source sonore, à partir d'une comparaison entre des caractéristiques respectives des signaux captés par les premier et deuxième capteurs ;
- déterminer (13) la première coordonnée spatiale de la source sonore sur la base de l'expression et de l'information supplémentaire.

(Figure 1)

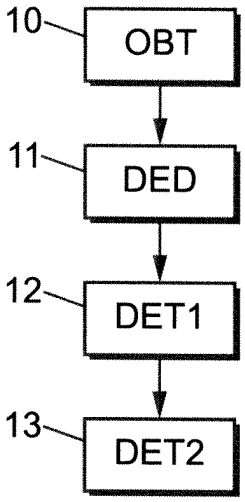


FIG. 1

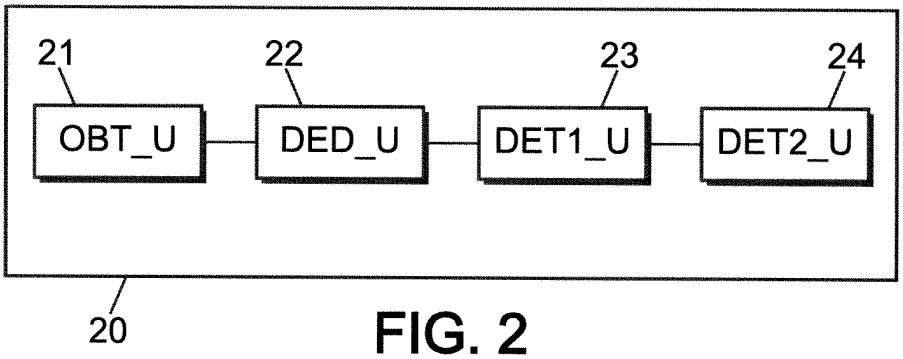


FIG. 2

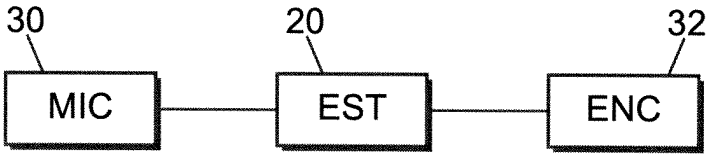
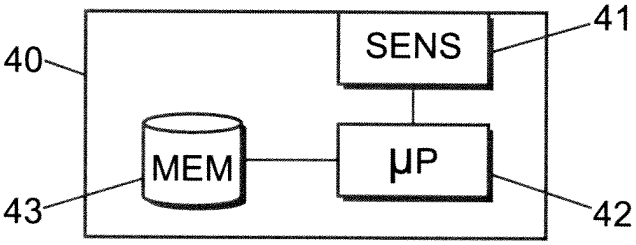


FIG. 3



308

FIG. 4

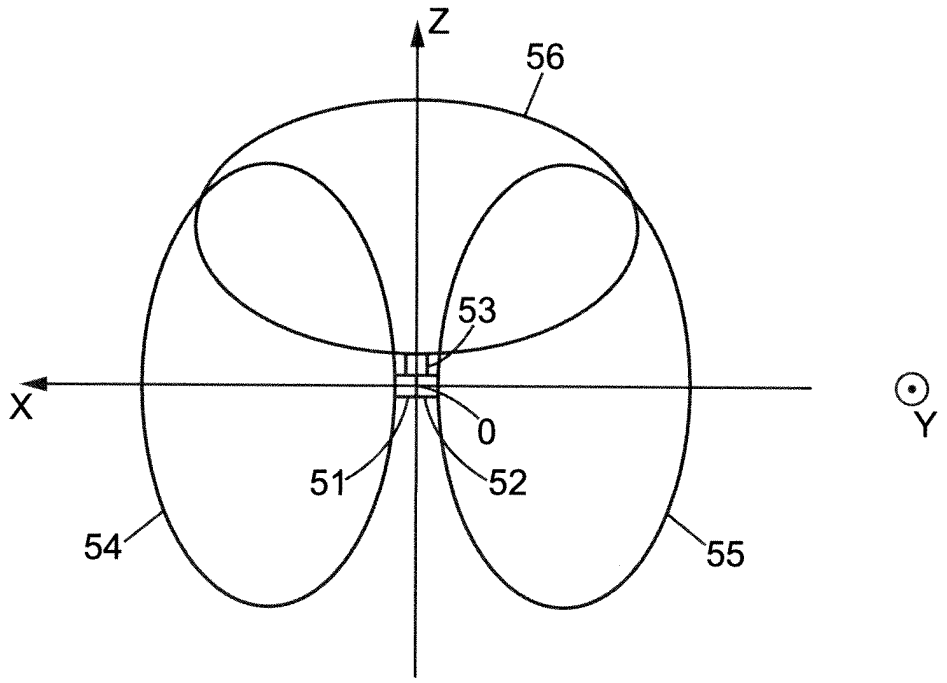


FIG. 5

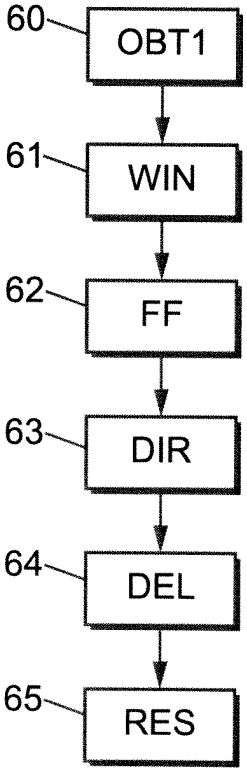


FIG. 6

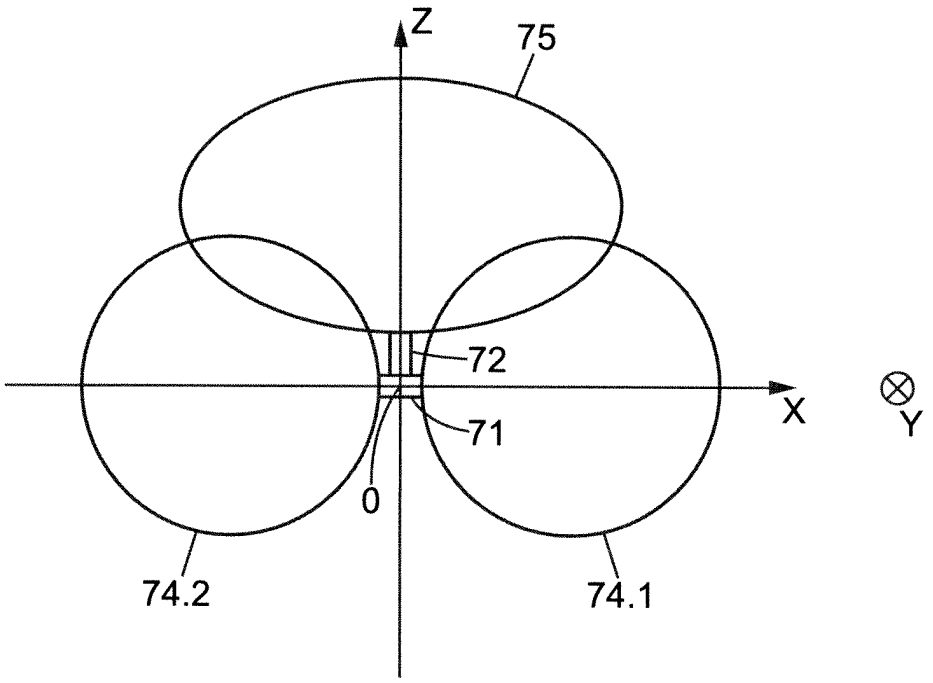


FIG. 7a

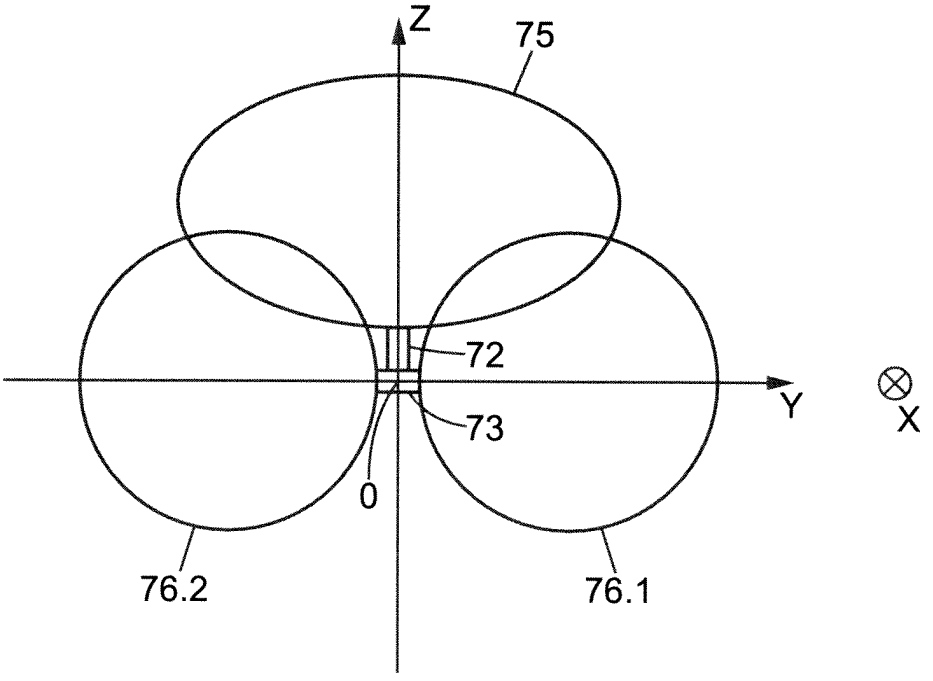


FIG. 7b

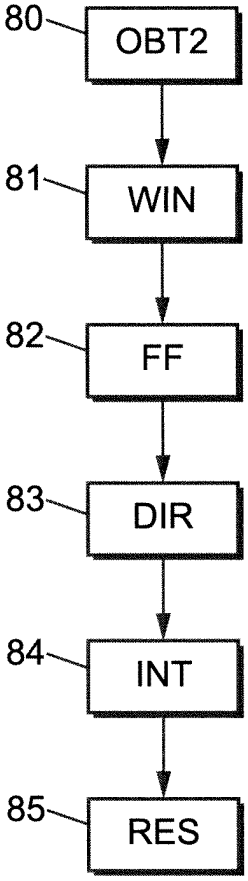


FIG. 8

Bibliographie

[AES, 1996] AES (1996). AES20-1996 : AES recommended practice for professional audio — subjective evaluation of loudspeakers.

[Algazi et al., 2001] Algazi, V., Duda, R., Thompson, D., and Avendano, C. (2001). The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2001 IEEE Workshop on*, page 99–102. IEEE.

[Asano et al., 1990] Asano, F., Suzuki, Y., and Sone, T. (1990). Role of spectral cues in median plane localization. *The Journal of the Acoustical Society of America*, 88(1) :159–168.

[Austrian Academy of Sciences, 2014] Austrian Academy of Sciences (2014). ARI HRTF database. `http://www.kfs.oeaw.ac.at/index.php?view=article&id=608&lang=en`.

[Ba et al., 2007] Ba, D. E., Florêncio, D., and Zhang, C. (2007). Enhanced mvdr beamforming for arrays of directional microphones. In *Multimedia and Expo (ICME), 2007 IEEE International Conference on*, page 1307–1310. IEEE.

[Bamford, 1995] Bamford, J. S. (1995). An analysis of ambisonic sound systems of first and second order. Master, University of Waterloo, Waterloo, Ontario, Canada.

[Beerends and Stemerdink, 1992] Beerends, J. G. and Stemerdink, J. A. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc*, 40(12) :963–978.

[Begault et al., 2001] Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc*, 49(10) :904–916.

[Berge and Barrett, 2010] Berge, S. and Barrett, N. (2010). A new method for b-format to binaural transcoding. In *Audio Engineering Society Conference : 40th International Conference : Spatial Audio —Sense the Sound of Space—*, Tokyo. AES.

[Berge and Barrett, 2014] Berge, S. and Barrett, N. (2014). Harpex. `http://harpex.net/`.

[Bili, 2014] Bili (2014). BiLi binaural listening. `http://www.bili-project.org/`.

[Blandin et al., 2011a] Blandin, C., Ozerov, A., Vincent, E., and others (2011a). Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Technical Report 7566, INRIA.

[Blandin et al., 2011b] Blandin, C., Vincent, E., and Ozerov, A. (2011b). Multi-source TDOA estimation using SNR-based angular spectra. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, page 2616–2619. IEEE.

[Blauert, 1983] Blauert, J. (1983). *Spatial Hearing - The Psychophisics of human sound localisation*. MIT Press, Cambridge, 1 edition.

[Blauert, 2013] Blauert, J. (2013). *The Technology of Binaural Listening*. Modern Acoustics and Signal Processing. Springer London, Limited.

[Blauert and Lindemann, 1986] Blauert, J. and Lindemann, W. (1986). Auditory spaciousness : Some further psychoacoustic analyses. *The Journal of the Acoustical Society of America*, 80(2) :533–542.

[Bluetooth Audio Video Working Group, 2002] Bluetooth Audio Video Working Group (2002). Bluetooth specification : Advanced audio distribution profile. *Bluetooth SIG Inc.*

[Breebaart, 2007] Breebaart, J. (2007). Analysis and synthesis of binaural parameters for efficient 3d audio rendering in MPEG surround. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, page 1878–1881. IEEE.

[Breebaart et al., 2008] Breebaart, J., Engdegaard, J., Falch, C., Hellmuth, O., Hilpert, J., Hoelzer, A., Koppens, J., Oomen, W., Resch, B., Schuijers, E., et al. (2008). Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding. In *Audio Engineering Society Convention 124th*, volume 7377, Amsterdam. AES.

[Breebaart et al., 2006] Breebaart, J., Herre, J., Jin, C., Kjörling, K., Koppens, J., Plogsties, J., and Villemoes, L. (2006). Multi-channel goes mobile : MPEG surround binaural rendering. In *Audio Engineering Society Conference : 29th International Conference : Audio for Mobile and Handheld Devices*, Seoul. AES.

[Breebaart and Kohlrausch, 2001] Breebaart, J. and Kohlrausch, A. (2001). The perceptual (ir) relevance of HRTF magnitude and phase spectra. In *Audio Engineering Society Convention 110th*, Amsterdam. AES.

[Breebaart et al., 2005] Breebaart, J., Par, S. v. d., Kohlrausch, A., and Schuijers, E. (2005). Parametric coding of stereo audio. *EURASIP J. Appl. Signal Process.*, 2005 :1305–1322.

[Bregman, 1994] Bregman, A. (1994). *Auditory Scene Analysis : The Perceptual Organization of Sound*. A Bradford book. Bradford Books.

[Bruneau, 1998] Bruneau, M. (1998). *Manuel d'acoustique fondamentale*. Etudes en mécanique des matériaux et des structures. Hermes.

[Burdic, 1991] Burdic, W. S. (1991). *Underwater acoustic system analysis*. Prentice Hall New Jersey.

[Butler and Humanski, 1992] Butler, R. A. and Humanski, R. A. (1992). Localization of sound in the vertical plane with and without high-frequency spectral cues. *Perception & psychophysics*, 51(2) :182–186.

[Byun, 2009] Byun, K. (2009). Digital audio effect system-on-a-chip based on embedded DSP core. *ETRI Journal*, 31(6) :732–740.

[Camacho and al, 2011] Camacho, R. and al (2011). Assessing the effect of 2D fingerprint filtering on ILP-Based structure-activity relationships toxicity studies in drug design. In *5th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2011)*, volume 93 of *Advances in Intelligent and Soft Computing*, pages 355–363. Springer Berlin Heidelberg.

[Carlile et al., 1997] Carlile, S., Leong, P., and Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. *Hearing research*, 114(1) :179–196.

[Cohen et al., 2010] Cohen, Y., Lewin, D., Kaplan, S., Sromin, A., and Simon, M. (2010). Digital speaker apparatus. US Patent App. 12/744,127.

[Committees, 2000] Committees, A. (2000). Press release. `http://www.aes.org/aeshc/docs/recording.technology.history/walkman2.html`.

[Dahl and Claesson, 1999] Dahl, M. and Claesson, I. (1999). Acoustic noise and echo cancelling with microphone array. *Vehicular Technology, IEEE Transactions on*, 48(5) :1518–1526.

[Damaske and Wagener, 1969] Damaske, P. and Wagener, B. (1969). Directional hearing tests by the aid of a dummy head. *Acta Acustica united with Acustica*, 21(1) :30–35.

[Daniel, 2011] Daniel, A. (2011). *Spatial Auditory Blurring and Applications to Multichannel Audio Coding*. PhD, Université Pierre et Marie Curie-Paris VI, Paris, France.

[Daniel et al., 2010] Daniel, A., McAdams, S., and Nicol, R. (2010). Multichannel audio coding based on minimum audible angles. In *Audio Engineering Society Conference : 40th International Conference : Spatial Audio —Sense the Sound of Space—*, Tokyo. AES.

[Daniel, 2001] Daniel, J. (2001). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD, Université Paris 6, Paris, France. text in french.

[Direction générale des relations culturelles, 2011] Direction générale des relations culturelles, scientifiques et techniques du Ministère des Affaires Etrangères, . (2011). Karlheinz stockhausen. `http://brahms.ircam.fr/composers/composer/3060/`.

[Dressler, 2000] Dressler, R. (2000). Dolby surround pro logic decoder principles of operation. *Dolby White paper*.

[du Gay et al., 2013] du Gay, P., Hall, S., Janes, L., Madsen, A., Mackay, H., and Negus, K. (2013). *Doing Cultural Studies : The Story of the Sony Walkman*. Culture, Media and Identities series. SAGE Publications.

[Du Moncel, 1887] Du Moncel, T. (1887). Le téléphone. In *Le téléphone*, Bibliothèque des merveilles, pages 117–127. Hachette, Paris, 5 edition.

[Duda et al., 1999] Duda, R. O., Avendano, C., and Algazi, V. R. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *Acoustics, Speech and Signal Processing (ICASSP), 1999 IEEE International Conference on*, volume 2, pages 965–968. IEEE.

[Dynes and Delgutte, 1992] Dynes, S. B. C. and Delgutte, B. (1992). Phase-locking of auditory-nerve discharges to sinusoidal electric stimulation of the cochlea. *Hearing Research*, 58(1) :79–90.

[EBU-tech, 1997] EBU-tech (1997). Assessment methods for the subjective evaluation of the quality of sound programme material – music.

[Enzner et al., 2011] Enzner, G., Krawczyk, M., Hoffmann, F. M., and Weinert, M. (2011). 3d reconstruction of HRTF-fields from 1d continuous measurements. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, page 157–160. IEEE.

[Faller, 2004] Faller, C. (2004). *Parametric coding of spatial audio*. PhD, EPFL, Lausanne, switzerland.

[Faller and Merimaa, 2004] Faller, C. and Merimaa, J. (2004). Source localization in complex listening situations : Selection of binaural cues based on interaural coherence. *The Journal of the Acoustical Society of America*, 116(5) :3075.

[Farina and Tronchin, 2005] Farina, A. and Tronchin, L. (2005). Measurements and reproduction of spatial sound characteristics of auditoria. *Acoustical science and technology*, 26(2) :193–199.

[Fastl and Zwicker, 2007] Fastl, H. and Zwicker, E. (2007). *Psychoacoustics : Facts and Models*. Springer series in information sciences. Springer.

[Faure, 2005] Faure, J. (2005). Evaluation de la synthèse binaural dynamique. Technical report, France Télécom R&D.

[Faure et al., 2007] Faure, J., Daniel, J., and Emerit, M. (2007). Optimisation d'une spatialisation sonore binaurale à partir d'un encodage multicanal. Patent WO 2007/101958 A2.

[Fletcher, 1934] Fletcher, H. (1934). Auditory perspective-basic requirements. *American Institute of Electrical Engineers, Transactions of the*, 53(1) :9–11.

[Fuldner et al., 2005] Fuldner, M., Dehe, A., and Lerch, R. (2005). Analytical analysis and finite element simulation of advanced membranes for silicon microphones. *IEEE Sensors Journal*, 5(5) :857–863.

[Furse, 2014] Furse, R. (2014). HOA technical notes - introduction to higher order ambisonics. `http://www.blueripplesound.com/hoa-introduction`.

[Furse and Malham, 2005] Furse, R. and Malham, D. (2005). Higher order ambisonics. `http://www.york.ac.uk/inst/mustech/3d_audio/secondor.html`.

[Gardner, 1997] Gardner, W. (1997). Three-dimensional audio using loudspeakers. Master, Massachusetts Institute of Technology, Massachusetts, USA.

[Geier et al., 2010] Geier, M., Ahrens, J., and Spors, S. (2010). Object-based audio reproduction and the audio scene description format. *Organised Sound*, 15(03) :219–227.

[Gerzon, 1973a] Gerzon, M. A. (1973a). Periphone (with height sound reproduction). In *Audio Engineering Society Convention 2ce*, volume 21, Munich. AES.

[Gerzon, 1973b] Gerzon, M. A. (1973b). Periphone (with height sound reproduction). In *Audio Engineering Society Convention 2ce*, volume 21, Munich. AES.

[Gerzon, 1975] Gerzon, M. A. (1975). The design of precisely coincident microphone arrays for stereo and surround sound. In *Audio Engineering Society Convention 50th*, volume 50, page 50, London. AES.

[Gerzon and Craven, 1977] Gerzon, M. A. and Craven, P. G. (1977). Coincident microphone simulation covering three dimensional space and yelding various directionall outputs. US Patent 4,042,779.

[Gerzon et al., 1999] Gerzon, M. A., Craven, P. G., Stuart, J. R., Law, M. J., and Wilson, R. J. (1999). The MLP lossless compression system. In *Audio Engineering Society Conference : 17th International Conference : High-Quality Audio Coding*, Florence. AES.

[Golub and Loan, 1996] Golub, G. H. and Loan, C. F. V. (1996). *Matrix computations*. JHU Press, 3th edition.

[Goodwin and Jot, 2008] Goodwin, M. and Jot, J. M. (2008). Spatial audio scene coding. In *Audio Engineering Society Convention 125th,*, San Francisco. AES.

[Goto et al., 2007] Goto, M., Iguchi, Y., Ono, K., Ando, A., Takeshi, F., Matsunaga, S., Yasuno, Y., Tanioka, K., and Tajima, T. (2007). High-performance condenser microphone with single-crystalline silicon diaphragm and backplate. *IEEE Sensors Journal*, 7(1) :4–10.

[Grantham, 1995] Grantham, D. W. (1995). Spatial hearing and related phenomena. *Hearing*, 6 :297–346.

[Grassi et al., 2003] Grassi, E., Tulsi, J., and Shamma, S. (2003). Measurement of head-related transfer functions based on the empirical transfer function estimate. In *Proceedings of the 2003 International Conference on Auditory Display*, page 119–122.

[Grey, 1977] Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5) :1270–1277.

[Groemer, 1996] Groemer, H. (1996). *Geometric Applications of Fourier Series and Spherical Harmonics*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

[Guillon, 2009] Guillon, P. (2009). *Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF*. Ph.D, Université du Maine, Le Mans, France. text in french.

[Guillon et al., 2012] Guillon, P., Zolfaghari, R., Epain, N., van Schaik, A., Jin, C. T., Hetherington, C., Thorpe, J., and Tew, A. (2012). Creating the sydney york morphological and acoustic recordings of ears database. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 461–466. IEEE.

[Guizart, 2011] Guizart, L. (2011). Mozart sur tous les écrans le 3 juin. *20 minutes*. `http://www.20minutes.fr/rennes/702383-rennes-mozart-tous-ecrans-3-juin`.

[Guski, 1997] Guski, R. (1997). Psychological methods for evaluating sound quality and assessing acoustic information. *Acta Acustica united with Acustica*, 83(5) :765–774.

[Gutiérrez Camarero and Moledero Dominguez, 2007] Gutiérrez Camarero, B. and Moledero Dominguez, I. (2007). Listening to music with headphones : An assessment of noise exposure and hearing damage. Headphone sound exposure and hearing, Aalborg University, Aalborg, Denmark.

[Hacihabiboglu et al., 2002] Hacihabiboglu, H., Gunel, B., and Murtagh, F. (2002). Wavelet-based spectral smoothing for head-related transfer function filter design. In *Audio Engineering Society Conference : 22nd International Conference : Virtual, Synthetic, and Entertainment Audio*, Espoo. AES, AES.

[Haidar, 2012] Haidar, N. (2012). Conception et développement d'un application audio stéréo sur plateforme de type iPhone. Master, ENSSAT Lannion, Lannion, France.

[Hall et al., 2008a] Hall, N., Okandan, M., Littrell, R., Bicen, B., and Degertekin, F. (2008a). Simulation of thin-film damping and thermal mechanical noise spectra for advanced micromachined microphone structures. *Journal of Microelectromechanical Systems*, 17(3) :688–697.

[Hall et al., 2008b] Hall, N., Okandan, M., Littrell, R., Serkland, D., Keeler, G., Peterson, K., Bicen, B., Garcia, C., and Degertekin, F. (2008b). Micromachined accelerometers with optical interferometric read-out and integrated electrostatic actuation. *Journal of Microelectromechanical Systems*, 17(1) :37–44.

[Hamasaki et al., 2005] Hamasaki, K., Hiyama, K., and Okumura, R. (2005). The 22.2 multichannel sound system and its application. In *Audio Engineering Society Convention 118th*, Barcelona. AES.

[Harder et al., 2013] Harder, S., Paulsen, R. R., Larsen, M., and others (2013). A three dimensional children head database for acoustical research and development. In *Proceedings of Meetings on Acoustics ICA*, volume 19, page 050013, Montreal. Acoustical Society of America.

[Hartung and Jonas, 1999] Hartung, K. and Jonas, B. (1999). Comparison of different methods for the interpolation of head-related transfer functions. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, volume AES 16th, Munich. AES.

[Herre et al., 2004] Herre, J., Faller, C., Ertel, C., Hilpert, J., Hoelzer, A., and Spenger, C. (2004). MP3 surround : Efficient and compatible coding of multi-channel audio. In *Audio Engineering Society Convention 116th*, Berlin. AES.

[Hiipakka, 2013] Hiipakka, M. (2013). Measuring pressure and particle velocity along the human ear canal. In *Proceedings of Meetings on Acoustics ICA*, volume 19, page 050019, Montreal. Acoustical Society of America.

[Hofman et al., 1998] Hofman, P. M., Van Riswick, J. G., and Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature neuroscience*, 1(5) :417–421.

[Huopaniemi and Smith, 1999] Huopaniemi, J. and Smith, J. (1999). Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, Munich. AES.

[Huopaniemi et al., 1999] Huopaniemi, J., Zacharov, N., and Karjalainen, M. (1999). Objective and subjective evaluation of head-related transfer function filter design. *J. Audio Eng. Soc*, 47(4) :218–239.

[IRCAM, 2014] IRCAM (2014). LISTEN HRTF database. `http://recherche.ircam.fr/equipes/salles/listen/index.html`.

[Ircam-Centre Pompidou, 2009] Ircam-Centre Pompidou, . (2009). Edgard varèse. `http://brahms.ircam.fr/edgard-varese`.

[Ircam-Centre Pompidou, 2012] Ircam-Centre Pompidou, . (2012). Iannis xenakis. `http://brahms.ircam.fr/iannis-xenakis`.

[ITU, 2003a] ITU (2003a). BS.1284-1 méthodes générales d'évaluation subjective de la qualité du son.

[ITU, 2003b] ITU (2003b). BS.1534-1 méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes de codage (MUSHRA).

[IUT, 1997] IUT (1997). BS 1116-1 methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.

[IUT, 2012] IUT (2012). BS.775-3 multichannel stereophonic sound system with and without accompanying picture.

[IUT, 2012] IUT (2012). BT.500 – méthodologie d'évaluation subjective de la qualité des images de télévision.

[Izumi et al., 2007] Izumi, Y., Ono, N., and Sagayama, S. (2007). Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2007 IEEE Workshop on*, volume 10, page 1000. IEEE.

[Jack and Thurlow, 1973] Jack, C. E. and Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the" ventriloquism" effect. *Perceptual and motor skills*, 37(3) :967–979.

[Jeelani, 2009] Jeelani, M. K. (2009). Integration and characterization of micromachined optical microphones. Master, Georgia Institute of Technology, Georgia, USA.

[Jessel, 1973] Jessel, M. (1973). *Acoustique théorique - Propagation et holophonie*. Masson, Paris, 1 edition.

[Johnston and Ferreira, 1992] Johnston, J. and Ferreira, A. (1992). Sum-difference stereo transform coding. In *Acoustics, Speech and Signal Processing (ICASSP), 1992 IEEE International Conference on*, volume 2, pages 569–572. IEEE.

[Jones and McManus, 1986] Jones, B. L. and McManus, P. R. (1986). Graphic scaling of qualitative terms. *SMPTE journal*, 95(11) :1166–1171.

[Joseph, 1980] Joseph, R. A. (1980). Hey, man ! new cassette player outclasses street people's 'Box.'. *The Wall Street Journal*, page 25.

[Jot et al., 1995] Jot, J.-M., Larcher, V., and Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. In *Audio Engineering Society Convention 98th*, Paris. AES.

[Katz, 2001] Katz, B. F. (2001). Boundary element method calculation of individual head-related transfer function. i. rigid model calculation. *The Journal of the Acoustical Society of America*, 110(5) :2440–2448.

[Keyrouz and Diepold, 2006] Keyrouz, F. and Diepold, K. (2006). An enhanced binaural 3D sound localization algorithm. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, pages 662 – 665.

[Keyrouz et al., 2006] Keyrouz, F., Diepold, K., and Dewilde, P. (2006). Robust 3d robotic sound localization using state-space hrtf inversion. In *Robotics and Biomimetics, 2006. ROBIO'06. IEEE International Conference on*, pages 245 – 250.

[Kistler and Wightman, 1992] Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91(3) :1637–1647.

[Knapp and Carter, 1976] Knapp, C. and Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4) :320–327.

[Kozamernik et al., 2007] Kozamernik, F., Marston, D., Mason, A., and Stoll, G. (2007). Ebu tests of multi-channel audio codecs. In *Audio Engineering Society Convention 122th*, Vienna. AES.

[Kuhn, 1977] Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(1) :157–167.

[Kuhn, 1987] Kuhn, G. F. (1987). Physical acoustics and measurements pertaining to directional hearing. In *Directional hearing*, pages 3–25. Springer.

[Kulkarni and Colburn, 1998] Kulkarni, A. and Colburn, H. S. (1998). Role of spectral detail in sound-source localization. *Nature*, 396(6713) :747–749.

[Kulkarni and Colburn, 2000] Kulkarni, A. and Colburn, H. S. (2000). Variability in the characterization of the headphone transfer-function. *The Journal of the Acoustical Society of America*, 107(2) :1071–1074.

[Kulkarni et al., 1995] Kulkarni, A., Isabelle, S., and Colburn, H. (1995). On the minimum-phase approximation of head-related transfer functions. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 1995 IEEE Workshop on*, pages 84–87. IEEE.

[Kulkarni et al., 1999] Kulkarni, A., Isabelle, S., and Colburn, H. (1999). Sensitivity of human subjects to head-related transfer-function phase spectra. *The Journal of the Acoustical Society of America*, 105(5) :2821–2840.

[Lane et al., 2010] Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9) :140–150.

[Lange, 2002] Lange, A. (2002). Histoire de la télévision. `http://histv2.free.fr/`.

[Lange, 2014] Lange, A. (2014). theatrophone.

[Langendijk and Bronkhorst, 2000] Langendijk, E. H. and Bronkhorst, A. W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *The Journal of the Acoustical Society of America*, 107 :528.

[Langendijk and Bronkhorst, 2002] Langendijk, E. H. and Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4) :1583–1596.

[Larcher, 2001] Larcher, V. (2001). *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD, Paris VI, Paris, France. text in french.

[Laster, 1983] Laster, D. (1983). Splendeurs et misères du théâtrophone. *Romantisme*, 13(41) :74–78.

[Lawless and Heymann, 1998] Lawless, H. T. and Heymann, H. (1998). Sensory evaluation of food. *Principles and practices*, pages 227–253.

[Le Bagousse, 2014] Le Bagousse, S. (2014). *Élaboration d'une méthode de test pour l'évaluation subjective de la qualité des sons spatialisés*. PhD, Université de Bretagne Occidentale, Brest, France. text in french.

[Le Bagousse et al., 2010] Le Bagousse, S., Colomes, C., and Paquier, M. (2010). Families of sound attributes for assessment of spatial audio. In *Audio Engineering Society Convention 129th*, San Francisco. AES.

[Le Bagousse et al., 2012] Le Bagousse, S., Paquier, M., and Colomes, C. (2012). Assessment of spatial audio quality based on sound attributes. In *Acoustics 2012*, Nantes, France.

[Le Bagousse and Paquier M., 2011] Le Bagousse, S. and Paquier M. (2011). Sound quality evaluation based on attributes - application to binaural contents. In *Audio Engineering Society Conference : 131th International Conference*, New York. AES.

[Letowski and Letowski, 2011] Letowski, T. and Letowski, S. (2011). Localization error : Accuracy and precision of auditory localization. *Advances in Sound Localization*, page 55–78.

[Lin and Sung, 2009] Lin, C.-Y. and Sung, C.-H. (2009). Multimedia audio dock. Patent US7593536 B2.

[Litovsky and Macmillan, 1994] Litovsky, R. Y. and Macmillan, N. A. (1994). Sound localization precision under conditions of the precedence effect : Effects of azimuth and standard stimuli. *The Journal of the Acoustical Society of America*, 96 :752.

[Ltd, 2014] Ltd, B. T. (2014). Theatrophone. `http://www.birchills.net/theatrophone/`.

[Lu et al., 2010] Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., and Campbell, A. T. (2010). The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, page 71–84. ACM.

[MacDonald, 2008] MacDonald, J. A. (2008). A localization algorithm based on head-related transfer functions. *The Journal of the Acoustical Society of America*, 123(6) :4290.

[Macpherson, 2009] Macpherson, E. A. (2009). Stimulus continuity is not necessary for the salience of dynamic sound localization cues. *The Journal of the Acoustical Society of America*, 125(4) :2691–2691.

[Magnoux and Lefort, 2014] Magnoux, J. and Lefort, A. (2014). Développement d'un démonstrateur audio 3d sur terminaux mobiles. Master, ENSSAT Lannion, Lannion, France. text in french.

[Majdak et al., 2007] Majdak, P., Balazs, P., and Laback, B. (2007). Multiple exponential sweep method for fast measurement of head related transfer functions. *J. Audio Eng. Soc.*

[Makous and Middlebrooks, 1990] Makous, J. C. and Middlebrooks, J. C. (1990). Two dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5) :2188–2200.

[Malham, 2008] Malham (2008). MTG : ambisonics home page. `http://www.york.ac.uk/inst/mustech/3d_audio/ambis2.htm`.

[Marschall and Chang, 2013] Marschall, M. and Chang, J. (2013). Sound-field reconstruction performance of a mixed-order ambisonics microphone array. In *Proceedings of Meetings on Acoustics ICA*, volume 19, page 055007, Montreal. Acoustical Society of America.

[McKeeg and McGrath, 1997] McKeeg, A. and McGrath, D. S. (1997). Using auralization techniques to render 5.1 surround to binaural and transaural playback. In *Audio Engineering Society Convention 102th*, Munich. AES.

[Mechanics, 1938] Mechanics, P. (1938). "big ears" listen for airplanes in mimic war raid over britain. *Popular Mechanics*, 70(6) :873.

[Merimaa and Pulkki, 2004] Merimaa, J. and Pulkki, V. (2004). Spatial impulse response rendering. In *Proceedings of 6th International Conference on Digital Audio Effects (DAFx-04)*, Naples, Italy.

[Middlebrooks, 1999] Middlebrooks, J. (1999). Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106 :1480.

[Middlebrooks and Green, 1991] Middlebrooks, J. C. and Green, D. M. (1991). Sound localization by human listeners. *Annual review of psychology*, 42(1) :135–159.

[Mills, 1958] Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4) :237–246.

[Minnaar et al., 1999] Minnaar, P., Christensen, F., Moller, H., Olesen, S. K., and Plogsties, J. (1999). Audibility of all-pass components in binaural synthesis. In *Audio Engineering Society Convention 106th*, Munich. AES.

[Minnaar et al., 2000] Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F., and Möller, H. (2000). The interaural time difference in binaural synthesis. In *Audio Engineering Society Convention 108th*, Paris. AES.

[Möller, 1992] Möller, H. (1992). Fundamentals of binaural technology. *Applied acoustics*, 36(3) :171–218.

[Moir, 1958] Moir, J. (1958). *High quality sound reproduction*. AEI Series. Chapman & Hall, London, 1st edition.

[Monaghan and Garlinghouse, 2013] Monaghan, C. and Garlinghouse, L. (2013). iTunes store sets new record with 25 billion songs sold. `http://www.apple.com/pr/library/2013/02/06iTunes-Store-Sets-New-Record-with-25-Billion-Songs-Sold.html`.

[Moreau, 2006] Moreau, S. (2006). *Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance*. PhD, Universite du maine, Le Mans, France. text in french.

[museum of retro technology, 2009] museum of retro technology, T. (2009). Acoustic radar. http ://www.aqpl43.dsl.pipex.com/MUSEUM/COMMS/ear/ear.htm.

[Nagra, 2014] Nagra (2014). Nagra FAQ. `http://www.nagraaudio.com/pro/pages/informationFaq.php`.

[Nam et al., 2008] Nam, J., Abel, J. S., and Smith III, J. O. (2008). A method for estimating interaural time difference for binaural synthesis. In *Audio Engineering Society Convention 125th*, San Francisco. AES.

[Nesta et al., 2009] Nesta, F., Svaizer, P., and Omologo, M. (2009). Cumulative state coherence transform for a robust two-channel multiple source localization. In *Independent Component Analysis and Signal Separation : 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings*, volume 5441, page 290–297. Springer.

[Nicol, 1999] Nicol, R. (1999). *Restitution sonore spatialisée sur une zone étendue : Application à la télépresence*. PhD, Université du Maine, Le Mans, France. text in french.

[Nicol, 2010] Nicol, R. (2010). *Représentation et perception des espaces auditifs virtuels*. HDR, Université du Maine - HDR Thesis, Le Mans, France. HDR Thesis, text in french.

[Nicol et al., 2014] Nicol, R., Gros, L., Colomes, C., Noisternig, M., Warusfel, O., Bahu, H., Katz, B. F., and Simon, L. S. (2014). A roadmap for assessing the quality of experience of 3d audio binaural rendering.

[Nujsop, 2013] Nujsop, C. M. (2013). Validation d'un prototype de création de contenus audio 3d pour les terminaux mobiles. Master, l'Ecole Nationale des Sciences Appliquées, Oujda, Morocco. text in french.

[Oldfield and Parker, 1984] Oldfield, S. R. and Parker, S. P. (1984). Acuity of sound localisation : a topography of auditory space. i. normal hearing conditions. *Perception*, 13(5) :581–600.

[Olson, 1978] Olson, H. F. (1978). *Modern sound reproduction*. R.E. Krieger Publishing Company, illustrated, reprint edition.

[Otani et al., 2010] Otani, M., Iwaya, Y., Suzuki, Y., and Itoh, K. (2010). Numerical analysis of HRTF spectral characteristics based on sound pressures on a pinna surface. In *Proceedings of 20th International Congress on Acoustics, ICA 2010*, page 1–8, Sydney.

[Painter and Spanias, 2000] Painter, T. and Spanias, A. (2000). Perceptual coding of digital audio. In *Proceedings of the IEEE*, volume 88, page 451–515.

[Palacino and Nicol, 2012] Palacino, J. and Nicol, R. (2012). Acquisition de données sonores spatialisées. Patent FR1260898.

[Palacino and Nicol, 2013a] Palacino, J. and Nicol, R. (2013a). Full 3D sound pick-up with a small microphone array : Prototype outline and preliminary assessment. In *Proceedings of the International Conference on Acoustics AIA/DAGA 2013*, Merano.

[Palacino and Nicol, 2013b] Palacino, J. and Nicol, R. (2013b). Spatial sound pick-up with a low number of microphones. In *Proceedings of Meetings on Acoustics ICA*, volume 19, page 055078, Montreal.

[Palacino and Nicol, 2013c] Palacino, J. and Nicol, R. (2013c). A surround microphone in your pocket. `http://www.acoustics.org/press/165th/4aSP2_Palacino.html`.

[Palacino and Nicol, 2014] Palacino, J. and Nicol, R. (2014). Des HRTF aux object-RTF : Système de prise de son 3d pour dispositifs nomades. In *CFA 2014*, Poitiers - France. CFA.

[Palacino et al., 2012] Palacino, J., Nicol, R., Emerit, M., and Gros, L. (2012). Perceptual assessment of binaural decoding of first-order ambisonics. In *Acoustics 2012*, Nantes, France.

[Pan, 1995] Pan, D. (1995). A tutorial on MPEG/audio compression. *IEEE multimedia*, 2(2) :60–74.

[Paulin et al., 1997] Paulin, P. G., Liem, C., Cornero, M., Nacabal, F., and Goossens, G. (1997). Embedded software in real-time signal processing systems : application and architecture trends. In *Proceedings of the IEEE*, volume 85, page 419–435.

[Pavel, 1983] Pavel, A. (1983). High fidelity stereophonic reproduction system. Patent US4412106 (A).

[Pernaux, 2003] Pernaux, J.-M. (2003). *Spatialisation du son par les techniques binaurales : application aux services de télécommunications*. PhD, I.N.P.G, Grenoble, France. text in french.

[Perrett and Noble, 1997] Perrett, S. and Noble, W. (1997). The contribution of head motion cues to localization of low-pass noise. *Perception & psychophysics*, 59(7) :1018–1026.

[Perrott, 1969] Perrott, D. R. (1969). Role of signal onset in sound localization. *The Journal of the Acoustical Society of America*, 45(2) :436–445.

[Perrott and Saberi, 1990] Perrott, D. R. and Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4) :1728–1731.

[Pike and Melchior, 2013] Pike, C. and Melchior, F. (2013). An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio. In *Audio Engineering Society Convention 134th*, Rome. AES.

[Poletti, 2000] Poletti, M. A. (2000). A unified theory of horizontal holographic sound systems. *J. Audio Eng. Soc*, 48(12) :1155–1182.

[Pollak, 1984] Pollak, A. (1984). Japan's stereo TV system. *New York Times*. `http://www.nytimes.com/1984/06/16/business/japan-s-stereo-tv-system.html`.

[Potel and Bruneau, 2006] Potel, C. and Bruneau, M. (2006). *Acoustique générale : équations différentielles et intégrales, solutions en milieux fluides et solides, applications*. TECHNOSUP. : Acoustique. Ellipses Marketing.

[Prager, 2012] Prager, J. (2012). l'interpretation acousmatique : fondements artistiques et techniques de l'interpretation des oeuvres acousmatiques en concert. *INA - GRM*.

[Preibisch-Effenberger, 1966] Preibisch-Effenberger, R. (1966). Zur methodik der richtungsaudiometrie : Prüfung der schallokalisationsfähigkeit durch elektroakustische verzögerungskette oder messungen im freien schallfeld ? *European Archives of Oto-Rhino-Laryngology*, 187(2) :588–592.

[press release, 2011] press release, T. I. (2011). Texas instruments spatial array IC simplifies audio soundstage design for multi-speaker portable products. *Reuters*.

[Pulkki, 2001] Pulkki, V. (2001). *Spatial sound generation and perception by amplitude panning techniques*. PhD, Helsinki University of Technology, Espoo, Finland.

[Pulkki, 2002] Pulkki, V. (2002). Compensating displacement of amplitude-panned virtual sources. In *Audio Engineering Society Conference : 22th International Conference : Virtual, Synthetic and Entertainment Audio*, page 186–195, Espoo. AES.

[Pulkki, 2006] Pulkki, V. (2006). Directional audio coding in spatial sound reproduction and stereo upmixing. In *Audio Engineering Society Conference : 28th International Conference : Future of sound technology - surround and beyond*, Pitea. AES.

[radiofrance.fr, 2014] radiofrance.fr (2014). Nouvoson. `http://nouvoson.radiofrance.fr/`.

[Rakerd and Hartmann, 1992] Rakerd, B. and Hartmann, W. M. (1992). Precedence effect with and without interaural differences - sound localization in three planes. *The Journal of the Acoustical Society of America*, 92(4) :2296–2296.

[Rayleigh, 1907] Rayleigh, L. (1907). XII. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74) :214–232.

[Reichinger et al., 2013] Reichinger, A., Majdak, P., Sablatnig, R., and Maierhofer, S. (2013). Evaluation of methods for optical 3-d scanning of human pinnas. In *3DTV-Conference, 2013 International Conference on*, pages 390–397. IEEE.

[Remple, 2003] Remple, T. B. (2003). USB on-the-go interface for portable devices. In *Consumer Electronics, 2003. ICCE. 2003 IEEE International Conference on*, pages 8–9. IEEE.

[Roth et al., 1980] Roth, G. L., Kochhar, R. K., and Hind, J. E. (1980). Interaural time differences : implications regarding the neurophysiology of sound localization. *The Journal of the Acoustical Society of America*, 68(6) :1643–1651.

[Royer et al., 1983] Royer, M., Holmen, J., Wurm, M., Aadland, O., and Glenn, M. (1983). ZnO on si integrated acoustic sensor. *Sensors and Actuators*, 4 :357–362.

[Rueff, 2010] Rueff, P. (2010). 3d radio - sytème chimera. `http://www.binaural.fr/binaural/?p=63`.

[Sanderson and Uzumeri, 1997] Sanderson, S. and Uzumeri, M. (1997). *Managing Product Families : The Case of the Sony Walkman*. McGraw-Hill international editions. Management and organization series. Irwin/McGraw-Hill.

[Sawada et al., 2007] Sawada, H., Araki, S., Mukai, R., and Makino, S. (2007). Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5) :1592–1604.

[Scavone et al., 2001] Scavone, G., Lakatos, S., Cook, P., and Harbke, C. (2001). Perceptual spaces for sound effects obtained with an interactive similarity rating program. In *Proceedings of International Symposium on Musical Acoustics*.

[Scheeper et al., 2003] Scheeper, P., Nordstrand, B., Gullov, J., Liu, B., Clausen, T., Midjord, L., and Storgaard-Larsen, T. (2003). A new measurement microphone based on MEMS technology. *Journal of Microelectromechanical Systems*, 12(6) :880–891.

[Scheeper et al., 1994] Scheeper, P. R., Van der Donk, A. G. H., Olthuis, W., and Bergveld, P. (1994). A review of silicon microphones. *Sensors and Actuators A : Physical*, 44(1) :1–11.

[Schmidt, 1986] Schmidt, R. O. (1986). Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34(3) :276–280.

[Schoenherr, 2005] Schoenherr, S. (2005). Recording technology history. `http://www.aes.org/aeshc/docs/recording.technology.history/notes.html`.

[Seppälä et al., 2006] Seppälä, E. T., Kirkeby, O., Kärkkäinen, A., Kärkkäinen, L., and Huttunen, T. (2006). Simulations of head related transfer functions in wideband acoustics. *The Journal of the Acoustical Society of America*, 119(5) :3430–3430.

[Smith, 2007] Smith, D. (2007). *Disney A to Z : The Official Encyclopedia*. Hyperion Books, New York, USA, 3th edition.

[Smith et al., 2014] Smith, E., Karp, H., and Wakabayashi, D. (2014). Apple in talks to buy beats electronics for $3.2 billion. *Wallstreet Journal*. `http://online.wsj.com/news/articles/SB10001424052702304431104579550392146532138`.

[Soulodre and Lavoie, 1999] Soulodre, G. A. and Lavoie, M. C. (1999). Subjective evaluation of large and small impairments in audio codecs. In *Audio Engineering Society Conference : 17th International Conference : High-Quality Audio Coding*, Florence. AES.

[Spiesberger, 2001] Spiesberger, J. L. (2001). Hyperbolic location errors due to insufficient numbers of receivers. *The Journal of the Acoustical Society of America*, 109(6) :3076–3079.

[Strybel and Perrott, 1984] Strybel, T. Z. and Perrott, D. R. (1984). Discrimination of relative distance in the auditory modality : The success and failure of the loudness discrimination hypothesis. *The Journal of the Acoustical Society of America*, 76(1) :318–320.

[Sunier, 1960] Sunier, J. (1960). *The story of stereo : 1881-*. Gernsback Library.

[Susini et al., 1999] Susini, P., McAdams, S., and Winsberg, S. (1999). A multidimensional technique for sound quality assessment. *Acta acustica united with Acustica*, 85(5) :650–656.

[Tajima et al., 2005] Tajima, T., Iguchi, Y., Goto, M., Ono, K., Ando, A., Tanioka, K., Takeshi, F., Matsunaga, S., and Yasuno, Y. (2005). An ultra-small high-performance silicon microphone using single-crystalline silicon. In *Consumer Electronics, 2005. ICCE. 2005 IEEE International Conference on*, pages 279–280.

[Thurlow and Runge, 1967] Thurlow, W. R. and Runge, P. S. (1967). Effect of induced head movements on localization of direction of sounds. *The Journal of the Acoustical Society of America*, 42(2) :480–488.

[University, 2014] University, N. (2014). Nagoya HRTFs database. `http://www.sp.m.is.nagoya-u.ac.jp/HRTF/database.html`.

[University, 2001] University, T. (2001). Tohoku university HRTFs database. `http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/`.

[Valin et al., 2007] Valin, J.-M., Michaud, F., and Rouat, J. (2007). Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3) :216–228.

[Van Wanrooij, 2005] Van Wanrooij, M. M. (2005). Relearning sound localization with a new ear. *Journal of Neuroscience*, 25(22) :5413–5424.

[Vanderlyn, 1978] Vanderlyn, P. V. (1978). In search of blumlein : The inventor incognito. *J. Audio Eng. Soc*, 26(9) :660–670.

[Vaslin, 2010] Vaslin, J.-M. (2010). Clément ader fut, aussi, l'inventeur de la stéréo, par jacques-marie vaslin. *Le monde*. `http://www.lemonde.fr/idees/article/2010/02/01/clement-ader-fut-aussi-l-inventeur-de-la-stereo-par-jacques-marie-vaslin_1299440_3232.html`.

[Viste and Evangelista, 2003] Viste, H. and Evangelista, G. (2003). On the use of spatial cues to improve binaural source separation. In *Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03)*, page 209–213, London, UK.

[Vogel et al., 2008] Vogel, I., Brug, J., Hosli, E., van der Ploeg, C., and Raat, H. (2008). MP3 players and hearing loss : adolescents' perceptions of loud music and hearing conservation. *The Journal of pediatrics*, 152(3) :400–404.

[Wahba, 1981] Wahba, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM Journal on Scientific Computing (SISC)*, 2 :5–16.

[Wallach, 1939] Wallach, H. (1939). On sound localization. *The Journal of the Acoustical Society of America*, 10(4) :270–274.

[Wallach, 1940] Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4) :339.

[Wasserman, 2004] Wasserman, L. (2004). définition 10.11. In *All of Statistics : A Concise Course in Statistical Inference*, Springer Texts in Statistics, page 461. Springer.

[Wenzel, 1995] Wenzel, E. (1995). The relative contribution of interaural time and magnitude cues to dynamic sound localization. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 1995 IEEE Workshop on*, pages 80–83. IEEE.

[Wenzel, 1999] Wenzel, E. M. (1999). Effect of increasing system latency on localization of virtual sounds. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*, Munich. AES.

[Wierstorf et al., 2011] Wierstorf, H., Geier, M., and Spors, S. (2011). A free database of head related impulse response measurements in the horizontal plane with multiple distances. In *Audio Engineering Society Convention 130th*, London. AES.

[Wightman and Kistler, 1992] Wightman, F. L. and Kistler, D. J. (1992). The dominant role of low frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3) :1648–1661.

[Wightman and Kistler, 1999] Wightman, F. L. and Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5) :2841–2853.

[Wolfson and Neumayr, 2008] Wolfson, R. and Neumayr, T. (2008). ABC, CBS, FOX & NBC offer incredible lineup of programming in stunning HD on the iTunes store. `http://www.apple.com/pr/library/2008/10/16ABC-CBS-FOX-NBC-Offer-Incredible-Lineup-of-Programming-in-Stunning-HD-on-the-iTunes-Store.html`.

[Woodworth et al., 1971] Woodworth, R., Schlosberg, H., Kling, J., and Riggs, L. (1971). *Woodworth & Schlosberg's Experimental psychology*. Holt, Rinehart and Winston.

[Zahorik, 2002] Zahorik, P. (2002). Auditory display of sound source distance. In *Proc. Int. Conf. on Auditory Display*, page 2–5.

[Zanesi and Gayou, 1983] Zanesi, C. and Gayou, E. (1983). Au coeur de l'acousmonium. `http://www.institut-national-audiovisuel.fr/sites/ina/medias/upload/grm/webmedia/2011/au_coeur_acousmonium/html5/co/Coeur_acousm1_montage.html`.

[Ziegelwanger et al., 2013] Ziegelwanger, H., Reichinger, A., and Majdak, P. (2013). Calculation of listener-specific head-related transfer functions : Effect of mesh quality. In *Proceedings of Meetings on Acoustics ICA*, volume 19, page 050017, Montreal. Acoustical Society of America.

[Zielinski et al., 2007] Zielinski, S., Brooks, P., and Rumsey, F. (2007). On the use of graphic scales in modern listening tests. In *Audio Engineering Society Convention 123th*, New York. AES.

[Zielinski et al., 2008] Zielinski, S., Rumsey, F., and Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. *J. Audio Eng. Soc*, 56(6) :427–451.

[Zone, 2012] Zone, R. (2012). *3-D Revolution : The History of Modern Stereoscopic Cinema*. UPCC book collections on Project MUSE. University Press of Kentucky, Kentuky.

[Zwislocki and Feldman, 2005] Zwislocki, J. and Feldman, R. (2005). Just noticeable differences in dichotic phase. *The Journal of the Acoustical Society of America*, 28(5) :860–864.

Curriculum Vitae

52, rue de Morlaix
22310 Plestin les Grèves
0686878430 - 0296461842
julianpalacino@hotmail.com

Julian PALACINO GARRIDO

Ingénieur Doctorant en Acoustique et traitement de signal

Né à Bogota, Colombie, le 30 octobre 1980
33 ans – Marié – Nationalité Française

COMPETENCES

Recherche bibliographique, veille technologique, réalisation de mesures et simulations acoustiques, rédaction de documents scientifiques, rédaction de brevets, animations scientifiques pour la valorisation de travaux, suivie de dossier, rédaction de rapports, relations et avec les clients, métrologie acoustique legale, veille normative, sensibilisation des clients et décideurs, définition des protocoles expérimentaux, développement d'indicateurs sous critères psychoacoustiques, réalisation de présentations, préparation de cours.

EXPÉRIENCES

2010-2013	**Ingénieur R&D** Orange Labs, *Lannion*	Doctorant «Outils de création de contenus audio spatialisés pour les terminaux mobiles»
2008-2010	**Ingénieur Acousticien** SPC Acoustique, *Montigny-lès-Metz*	Chargé d'études en acoustique environnementale, industrielle et du bâtiment.
Avril-Août 2008	**Stage de fin d'études** CSTB, *Grenoble*	Exposition sonore du cycliste lors de son déplacement en milieu urbain.
Avril-Juin 2007	**Stage de recherche** LAUM, *Le Mans*	Acoustique urbaine, mesure du champ acoustique dans une rue et influence des parois.
Février-Mai 2006	**Projet de licence acoustique et mécanique** Université du Maine, ODES, ESEO	Fabrication d'un dispositif pour la mesure des niveaux acoustiques émis par des baladeurs (tête artificielle).
Avril-Juin 2005	**Stage de technicien acousticien** ISVR, *Southampton*	Etude et localisation des impulsions produites par des organismes vivants sur les côtes anglaises.

ENSEIGNEMENT

2010-2011 **Enseignant Vacataire -** Université de Rennes 1, Electronique DUT Réseaux et Télécom, *Lannion*

2013 **Enseignant Vacataire -** Université de Rennes 1, Licence pro CIAN, *St. Brieuc*

FORMATION

2007-2008 **Master 2** Professionnel, Acoustique des Transports. Université du Maine (*Le Mans, France*)

2006-2007 **Master 1** Mécanique et Acoustique, Option Acoustique. Université du Maine (*Le Mans, France*)

2005-2006 **Licence** Mécanique, Option Acoustique. Université du Maine (*Le Mans, France*)

2003-2005 **DEUST** Vibrations, Acoustique et Signal, Option Parole et Sonorisation. Université du Maine (*Le Mans, France*)

2002-2003 1ère Année **DUPM** Pédagogie Musicale, Electroacoustique. Conservatoire de Perpignan (*Perpignan, France*)

2000-2001 **Equivalent DEUG** Études ingénierie du Son. Pontificia Universidad Javeriana (*Bogota, Colombie*)

Julian PALACINO GARRIDO

52, rue de Morlaix
22310 Plestin les Grèves
0686878430 - 0296461842
julianpalacino@hotmail.com

Ingénieur Doctorant
en Acoustique et traitement de signal

Né à Bogota, Colombie, le 30 octobre 1980
33 ans – Marié – Nationalité Française

Langues

Espagnol: Langue Maternelle.

Français: Bilingue.

Anglais: Avancé (915/990 TOEIC obténu en 02/2014. Stage de 3 mois en Angleterre, séjours au Canada et USA).

Italien: Basique (Plusieurs séjours en Italie)

Informatique

Logiciels scientifiques et langages de programmation.

C, Matlab, VBA, LaTex, Labview, Max, Mithra SIG, CadnaA, Catt-Acoustic, Ease, Acoubat, Logiciels et systèmes de metrologie 01 dB et B&K.

Bureautique, Infographie et Audio sur PC et MAC.

Maîtrise en création multimédia avec Suite Macromedia, Suite Adobe, AutoCad, Digital Performance, Protools, Reaper, Live, Office, développement de sites web.

Publications

J. Palacino, *Outils de spatialisation sonore pour terminaux mobiles : microphone 3D pour une utilisation nomade en exploitant la localisation des sources en vue d'une description «objet»*, PhD, Université du Maine, Le Mans, France, 2014.

J. Palacino et R. Nicol, « Acquisition de données sonores spatialisées », CFA, Poitiers, France, 2014.

J. Palacino et R. Nicol, « Full 3D sound pick-up with a small microphone array: Prototype outline and preliminary assessment », in Proceedings of the International Conference on Acoustics, Merano, 2013.

J. Palacino et R. Nicol, « Spatial sound pick-up with a low number of microphones », in Proceedings of Meetings on Acoustics, Montreal, 2013, vol. 19, p. 055078.

J. Palacino et R. Nicol, « A Surround Microphone in Your Pocket », ASA Lay Language Papers 165th Acoustical Society of America Meeting. [En ligne]. Disponible sur: http://www.acoustics.org/press/165th/4aSP2_Palacino.html.

J. Palacino and R. Nicol, «ACQUISITION DE DONNÉES SONORES SPATIALISÉES,» FR1260898 / CIB: H04S7/00 H04R5/00 G01S5/22, 16-Nov-2012.

J. Palacino, R. Nicol, M. Emerit, et L. Gros, « Perceptual assessment of binaural decoding of first-order ambisonics», in Acoustics 2012, Nantes, France, 2012.

J. Defrance et J. Palacino, « Auscultation acoustique des aménagements cyclables en milieu urbain », in 10eme Conges Français d'Acoustique, Lyon, 2010.

J. Defrance, J. Palacino, M. Baulac, and others, "Acoustical assessment of cycle paths in urban areas," in Euronoise 2009, Edinburgh, Scotland, 2009.

Informations Complémentaires

- Pratique musicale, basse électrique.
- Passionné par le bricolage, la mécanique, la photographie argentique et numérique, la danse et la plongée.
- Titulaire des brevets de plongeur niveau 2 CMAS et Advanced Scuba Diving NAUI.
- Titulaire du permis de conduire A, B.